NAVAL RESEARCH LOGISTICS QUARTERLY

JUNE 1976
VOL. 23, NO. 2

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics a will publish research and expository papers, including those in certain areas of mathematics, statistics, and econom relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, Ju September, and December and can be purchased from the Superintendent of Documents, U.S. Government Print Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulati P-35 (Revised 1-74).

REDUNDANT SPARES ALLOCATION TO REDUCE RELIABILITY COSTS*

Leonard Shaw and Sharad G. Sinkar*

Polytechnic Institute of New York
Department of Electrical Engineering and Electrophysics
Brooklyn, New York

ABSTRACT

The problem considered here is the optimal selection of the inventory of spares for a system built from two kinds of modules, the larger of which can be connected so it performs the role of the smaller one. The optimal inventory is the least costly one which achieves a specified probability that the spares will not be exhausted over the design lifetime. For some costs and failure rates it is most economical to use the larger module for both roles, due to the resulting increase in flexibility in the deployment of a single type of spare module.

Both analytical and simulation methods have been used to study this problem.

. INTRODUCTION

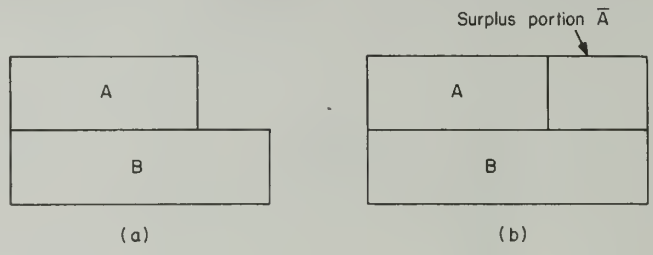
An optimum spares inventory is one which minimizes a combination of cost, weight, etc., while chieving a specified high probability that the spares, procured simultaneously at the time of equip- ıent acquisition, will not be exhausted during the planned lifetime of the equipment. This paper onsiders several approaches to the design of optimum spares inventories for a two module system in vhich one module is a subset of the other.

In many modern complex systems one of the modules may be a submodule of another module. 'he possible candidates could be amplifiers and an amplifier-power supply, or amplifiers and amplifier- etectors. Furthermore, it might be advantageous to artificially create this kind of redundant situation, .g., including "read" and "write" circuitry on every module in a memory unit when one read-write ircuit can actually serve several memory modules. For some combination of costs, failure rates, etc., : may be efficient to use only the larger of two such modules. (Alternatively, even if both kinds are used ı the original equipment, it may be efficient to stock spares of only the larger type.) The cost of intro- ucing and stocking a separate module can be saved, and the inventory of a single module with multiple ses can be more flexible.

A Discard at Failure Maintenance philosophy (D.A.F.M.) is assumed. That is, the maintenance ctivity consists of locating a failed module and replacing it with a new one obtained from the spares ıventory. In some of the approaches considered here we do allow the possibility of a module which is FAILED" as a bigger module but which can be used, without any repair, as the smaller one.

*This work partially supported by the Office of Naval Research under Contract No. N0014-67-A-0438-0013.

In short, we consider the bigger module B comprised of two submodules, A (smaller module) and $\bar{A}$. It is assumed that failure of B is due to the failure of either A or $\bar{A}$ but not due to the failure of both. If a B module fails due to failure of $\bar{A}$, then the module can be used as the smaller module A. These cases are depicted in Figure 1.



(a) A AND B ARE SEPARATE MODULES
(b) B IS USED AS A WITH Ā AS SURPLUS PORTION

FIGURE 1. Two realizations of a two module system.

Several steps are followed to evaluate and compare different spares inventories. A module reliability model must be combined with the repair discipline and inventory description in order to compute the spares adequacy. The cost of a particular inventory is based on a model of costs for fabrication, acquisition and storage of separate modules.

Six different inventory and repair models are considered.* The conventional stocking of both kinds of modules, and use of each in only its original function, is contrasted with several disciplines in which B modules may be used in the simpler A function. In one case all modules in the original equipment and in the spares inventory are the larger B type. In another case both kinds of spares are stocked, and use of B's in A functions is allowed. These last two cases are further subdivided into either idealized subcases which can be solved analytically, or more practical versions which can only be evaluated via simulation.

The steps for formulating and evaluating this two module spares adequacy problem will now be described in more detail.

2. RELIABILITY AND COST MODELS

The reliability model considered here is based on Reference 4. Appendix A describes how the failure rates λ_i (for the constant hazard failure model used here) are calculated for the circuits considered as modules.

One advantage of the spares redundancy approach is that a failed large module B might still function perfectly in an A task. Using this approach it is necessary to compute the probability of such an event. The constant hazard failure models and assumed independence of failures in A and $\bar{A}$ portions of a B module make the computation quite simple, as follows.

The probability that submodule A has failed given that B has failed is equivalent to the probability that the exponentially distributed time-to-A failure is less than the independent time-to-$\bar{A}$ failure. An easy calculation [Reference 5] yields

*We are grateful to a referee whose comments suggested the addition of some of these cases.

1) $$P[A \text{ failed} \mid B \text{ failed}] = \frac{\lambda_1}{\lambda_2},$$

where

λ_1 = failure rate of submodule A

λ_2 = failure rate of module B

$\lambda_2 - \lambda_1$ = failure rate of submodule $\bar{A}$.

Similarly, the probability of submodule A being good when B fails is

$$P[A \text{ good} \mid B \text{ failed}] = 1 - \frac{\lambda_1}{\lambda_2}.$$

The model used to describe the cost of equipment and spares acquisition, along with storage costs, s developed in Reference 2. The result is summarized in the following total cost expression:

2) $$C_T = n(CLI + NL \cdot CSC) + \sum_{i=1}^{n} C_i[n_i + NE \cdot \nu_i],$$

n which

n = Number of different kinds of modules

n_i = Number of i-modules acquired as spares

ν_i = Number of i-modules per equipment

NL = Design lifetime in years

NE = Number of equipments

CLI = Cost of introducing a line item into inventory

CSC = Shelf cost per year

C_i = Purchase cost for module i.

The cost model for a single module is further decomposed into

3) $$C_i = C_a N_i + C_b P_i + C_c,$$

where

C_i = Cost of the ith module type

N_i = Number of components in the ith module type

P_i = Number of external pins required for ith module type.

In the cases where a B module is used for the A function, the pin configuration must be compatible. However, the unequal values of P_i for the two functions will refer to the number of pins whose failure would interrupt proper functioning. The pin dependent cost includes both fabrication costs corresponding to forming electrical connections to the pins as well as the cost of the mechanical connector. The two modules must have identical connectors if an A function may accept a B module. If the electrical fabrication costs dominate then (3) is appropriate. If the mechanical connector cost dominates then P_B should be used for P_i in (3), for *both* module costs.

In the optimizations and tradeoffs discussed below, all of the above parameters will be fixed except for n and n_i. Table 1 summarizes the parameter values (taken from Reference 2) which were used in the following examples.

TABLE 1. *Module Parameter Values*

Constant	Description	Value
i	=1 for module A, =2 for module B	
C_a	Component cost factor	\$2
C_b	Interconnect cost factor	\$2
C_c	Package cost factor	\$100
NE	Number of equipments	20
CLI	Introduction cost of new line item	\$100
CSC	Shelf cost per year	\$12
NL	Planned years of equipment life	10 years
ν_i	Number of i modules/equipment	$1(i=1, 2)$
N_A	Number of components in A	35
N_B	Number of components in B	50
P_A	Number of pins for A	10
P_B	Number of pins for B	16

The cost model in (2) is specifically for physically small integrated circuits for which the shelf costs depend on the number of bins needed to store separate module types, but not on the number of modules stored. Moreover, this model assumes for simplicity that all modules are available for use in all systems. If the individual systems are widely dispersed, this central stocking will result in increased turnaround time which is not explicitly accounted for here. Modification to other stocking arrangements would be straightforward.

3. INVENTORY OPTIMIZATION

Six different kinds of inventory optimizations have been evaluated and compared. (See Table 2.)

TABLE 2. *Comparison of Replacement Disciplines*

Case	
1. Store A and B	Each used only for its own function.
2. Store B only, and only B in original equipment.	If possible, partially failed B exchanged with completely good B which was in A function. Otherwise store partially failed B and replace with completely good spare B. Preference in replacing failed A functions is i) stored partially good B, or ii) completely good B spare.
3. Store B only, and only B in original equipment.	Same as 2 except no retrieval of completely good B from A function.
4. Store A and B	Same as 2 except use up spare A's first for replacing failed A function.
5. Store A and B	Same as 4 except no retrieval of completely good B from A function.
6. Store A and B	Discard partially good B's. Completely good B may be used in A function when A spares are exhausted.

The first finds the best (or approximately so) stocks of both A and B modules to minimize costs while achieving an acceptable probability that the spares will be adequate for the designed equipment lifetime. In that case each module is used only in its originally intended function. The other approaches allow substitution of B modules for the simpler A function, using different procedures for allocating partially good and completely good spares.

In the first case, using both A and B modules, the probability that the inventory of n_j of a module of type j is adequate is

$$P_j(n_j) = \sum_{k=0}^{nj} e^{-\theta_j} \frac{(\theta_j)^k}{k!}, \tag{4}$$

where θ_j is the expected number of failures of module j during the lifetime of NL years. The failure rate λ_j in percent/thousand hours computed as in Appendix A can be used to compute θ_j according to

$$\theta_j = \lambda_j \cdot NL \times NE = 0.0876. \tag{5}$$

This formula assumes, for simplicity, that the λ_j are known constants and that the modules will be operated continuously, 365 days a year. Moreover, it is assumed that modules do not fail while in storage, that good modules are never erroneously discarded, and that no modules suffer secondary failures as a result of the failure of other equipment, repairman errors, etc.

The total inventory will be adequate only if the stock of each type of module is adequate, so the adequacy P_T is defined by

$$P_T = P_A(n_A)P_B(n_B). \tag{6}$$

While the optimal inventory to minimize cost while achieving a specified P_T can, in principle, be computed exactly via dynamic programming, an alternate and more feasible approach is taken here to get a good approximation of the optimal inventory. This approximate optimization finds the best inventory from among an incomplete set of undominated allocations.

A policy is said to be "undominated" if, for any policy that yields a better system adequacy, more of at least one resource is required. In our case, of course, there is only one resource, i.e., cost. The complete family of undominated allocations is the set consisting of all undominated allocations. The solution to any problem must be a member of the complete family. An incomplete family does not consist of all undominated allocations, and it is quite possible that the solutions to a particular problem may not be a member of the incomplete family, though each member of the incomplete family is undominated.

To explain this further, consider two adequacy levels from the computer solution (Table 3). For total cost of \$11,460, the solution is $A = 5$, $B = 7$ and the adequacy level = 0.993; and for the total cost of \$11,692, the solution is $A = 5$, $B = 8$ and the adequacy level = 0.997. Suppose now that the total funds available are \$11,550. Then from our incomplete family, the best solution is the one corresponding to the total cost of \$11,460. In reality, however, the optimum solution could be $A = 4$, $B = 8$ and the adequacy level = 0.994. This not being a member of the incomplete family, we have to settle for a nearly optimum solution. In practice, however, the members of the incomplete family

are so close to each other that the difference between the optimum solution and approximately optimum solution is negligible.

TABLE 3

ANOS	ANCC	TMW	COMP	PN
219.0	103.0	5.6	35.0	10.0
437.0	229.0	8.3	50.0	16.0

NE = 20.0 ANL = 10.0
Lambda(1) = 0.072
Lambda(2) = 0.148

TWO MODULE TYPES—Case 1

Total cost	Number of *A*	Number of *B*	Adequacy level
$8,886.00	0.0	0.0	0.021
9,118.00	0.0	1.0	0.077
9,308.00	1.0	1.0	0.173
9,540.00	1.0	2.0	0.335
9,730.00	2.0	2.0	0.452
9,962.00	2.0	3.0	0.641
10,194.00	2.0	4.0	0.762
10,384.00	3.0	4.0	0.845
10,616.00	3.0	5.0	0.915
10,806.00	4.0	5.0	0.943
11,038.00	4.0	6.0	0.974
11,270.00	4.0	7.0	0.986
11,460.00	5.0	7.0	0.993
11,692.00	5.0	8.0	0.997
11,882.00	6.0	8.0	0.998
12,114.00	6.0	9.0	0.999

ONE MODULE TYPE—Case 2

Total cost	Number of *B*	Adequacy level
$9,503.00	0.0	0.081
9,735.00	1.0	0.284
9,967.00	2.0	0.540
10,199.00	3.0	0.754
10,431.00	4.0	0.889
10,663.00	5.0	0.957
10,895.00	6.0	0.985
11,127.00	7.0	0.996
11,359.00	8.0	0.999
11,591.00	9.0	1.000
11,823.00	10.0	1.000
12,055.00	11.0	1.000
12,287.00	12.0	1.000
12,519.00	13.0	1.000
12,751.00	14.0	1.000

We generate an incomplete family of undominated allocations by constructing successively larger redundancy allocations. This is achieved by adding one module at a time. The module we add will be the one that provides greater improvement in the system reliability per 100 dollars spent. (While it

is not in the range of usable reliabilities, the inventory of a single A and no B's is clearly undominated but does not appear in the incomplete family generated this way.)

It can be shown [1, p. 167] that each allocation obtained in the above procedure is undominated if $\log P_T$ is a concave function. This is satisfied because the failure rates are all constant [1, p. 175].

The computational steps for such a procedure are as follows:

(1) Compute adequacy of the system with zero spares initially allocated.

(2) Compute for each module one increment in the logs of the module adequacy per 100 dollars for one added spare.

$$[\text{Log } P_T(n_i + 1) - \text{Log } P_T(n_i)] \times 100/C_i = \Delta.$$

(3) Choose that module which maximizes Δ and allocate one spare to that module type.

(4) If both the module types give the same increment in the adequacy, then select the module of smaller cost.

(5) Repeat the procedure from (2) to (4) using the spares policy previously generated until the maximum allowable cost is reached.

The second and third cases (see Table 2), which use only B modules in original equipment and as spares, are differentiated according to how they use completely good and partially good spares. *The second approach* assumes that if an A-function module fails when there is no partially failed B, that failed A is replaced by a new B. Subsequently that B is replaced by a partially failed B as soon as one is available. (The completely good B is then returned to the spare inventory.) Moreover, a partially failed B will be interchanged with a completely good B which was operating in an A function, if any are being used in this way. These assumptions are somewhat unrealistic, but make the adequacy of a particular inventory easier to examine analytically. If such an interchange of partially and completely good B modules is not allowed, the analysis become intractable. (That more realistic situation, without such interchanges, comprises Case 3 which is handled via simulation.)

Another assumption made here is that this exchange of good B and partially failed B does not interrupt equipment operation appreciably and incurs no additional maintenance cost. Furthermore the $\bar{A}$ part of a B module does not fail while the module is operating in an A function.

Let

KB = total # of B spares

CB = # of complete failures of B modules

PB = # of partial failures of B modules

NA = # of failures of A-function modules

NE = # of equipments.

While the adequacy of a particular inventory will depend on the order in which various kinds of failures occur, it is possible to get simple expressions for upper and lower bounds on the adequacy probability. The upper bound assumes that for any given number of partial B failures, all occur before any A failures. In this way there are $(KB + NE - CB)$ spares for replacing partial failures and the upper bound P_u (adequacy) is given by

$$P_u(\text{adequacy}) = \sum_{l=0}^{KB} \{P(CB = l) \sum_{n=0}^{KB-l} \left[P(NA = n) \sum_{m=0}^{KB-l+NE} P(PB = m) \right] \}. \tag{7a}$$

The term NE appears in the above expression because there are NE good B modules which were originally used as A modules when the equipments were procured. The effective number of maximum allowable partial failures would, therefore, be the total number of spares plus number of equipments procured.

The lower bound P_l can be found similarly by assuming the worst failure ordering in which, for any given number of A failures, all precede any partial B failures. In that case only $(NE+KB-CB-NA)$ partial B failures can be replaced, and the resulting probability is

$$\text{(7b)} \qquad P_l\,(\text{adequacy}) = \sum_{l=0}^{KB} \left\{ P(CB = l) \sum_{n=0}^{KB-l} \left[P(NA = n) \sum_{m=0}^{NE+KB-l-n} P(PB = m) \right] \right\}.$$

In both of these extreme orderings, or any other one, an A failure must be replaced from stock. There will always be a total of $KB - CB - NA$ spares, either all completely good or a mix of completely and partially good ones. Partially good B's are stored when $PB > NE$, in which case a partial failure is replaced from the spares rather than by a completely good B from an A function. Furthermore, in this pessimistic ordering, partially failed B modules are never used in A functions.

The actual adequacy probability must be between P_u and P_l. Fortuitously, these two bounds yield identical numerical values in the examples presented below. This tightness of the bounds results whenever the number of equipments is fairly large, so that even the minimum number NE of partial B spares far exceeds the KB spares which must be shared by A and complete B failures.

The terms in the foregoing adequacy expressions can be further explained as follows. An A-function failure and a complete B failure both correspond to the failure of the A portion of a B module. While these events need not have been separated in (7), those expressions define A failures when the failed module is in use as an A function. With this in mind, the NA and CB probabilities in (7) are both of form (4) and (5) with $\lambda_j = \lambda_1$. Similarly, the number of partial failures follows the Poisson law with $\lambda = (\lambda_2 - \lambda_1)$.

The best inventory using these failure probabilities is simply the smallest KB which achieves the desired P (adequacy). Since there is only one module type, there are no tradeoffs to be considered and the adequacy probability increases monotonically with the number of spares.

The third inventory design takes a more realistic approach to the use of a single module type. Partially failed B modules are stored and replaced by completely good ones. If a partially failed B is available, it is used to replace each failed A module. When no partially failed B's are available, a completely good one is used to replace a failed A-function module, but that module is kept in use until its A portion fails. The shifting of completely and partially good modules used in the previous discipline is not allowed here.

This prohibition against extra interchanges makes the adequacy even more dependent on the order of A and B failures. Upper and lower bounds on the adequacy can be computed from expressions which differ from (7a) and (7b) only by the deletion of the NE which appears in the upper limit of a summation. These bounds were significantly different in the numerical example considered below, so a Monte Carlo simulation was employed to get a more precise value for the adequacy probability. The simulation approach takes advantage of the control variable method for variance reduction. To that end, both the analyzable interchange discipline (with tight upper bound (7a)) and the nonretrieval discipline are simulated simultaneously using the same pseudorandom numbers. Comparison of analytical and

simulated adequacy for the interchange discipline provides correction terms for the simulation values of the nonretrieval discipline.

Complete flow charts and FORTRAN programs for all of these calculations are available in Reference 7.

Cases 4 and 5 are intermediate between the conventional *Case 1* and the single module *Cases 2 and 3*. Here the original equipment uses *B*'s for *B* functions and *A*'s for *A* functions, but use of completely or partially good *B*'s as replacements in *A* functions is allowed. *Case 4* allows retrieval of completely good *B*'s from *A*-function duty when partially failed *B*'s become available. While this unrealistic interchange again reduces dependence of adequacy on the order of module failures (as in Case 2), the adequacy iteration used in Case 1 to generate incomplete, undominated inventories is not applicable here. That efficient technique requires stochastic independence of the random failure and replacement processes for the separate module functions. The sharing of *B* spares between both functions violates that assumption.

Case 4 has, therefore, been evaluated by exhaustive calculation of the cost and adequacy of all possible inventories for which the number of spares is less than a fixed limit. These inventories were ordered according to cost, and dominated inventories were culled out of the sequence to get a complete family of undominated policies. (This brute force approach may be feasible only for simple illustratives like the one considered here.)

In this case the adequacy probability for an inventory of *KB B* spaces and *KA A* spares can be bounded in a manner similar to the technique used for Case 2. The upper bound results when, for any given number of partial *B* failures, all occur before any *A* failures:

$$P_u(\text{adequacy}) = \sum_{l=0}^{KB} \left\{ P(CB=l) \sum_{n=0}^{KA+KB-l} P(NA=n) \sum_{m=0}^{KB-l} P(PB=m) \right\}, \tag{8a}$$

which is quite similar to Equation (7a) for Case 2.

A lower bound results when, for any given number of *A* function failures, all precede any partial *B* failures. In this situation only $[KB-CB-Max\ (0, NA-KA-NE)]$ spares are available for partial *B* replacements:

$$P_l = \sum_{l=0}^{KB} \left\{ P(CB=l) \sum_{n=0}^{KA+KB-l} \left[P(NA=n) \sum_{m=0}^{KB-l-Max(0,\, n-KA-NE)} P(PB=m) \right] \right\}, \tag{8b}$$

which is similar to Equation (7b) for Case 2.

Case 5 is the more realistic version of 4, in which the retrieval of a completely good *B* module from an *A* function is not allowed. This situation makes adequacy even more dependent on the order of *A*-function and *B*-function failures. Equation (8a) gives an upper bound on this adequacy probability. A lower bound is provided by (8b) with the *NE* deleted from the expression in the upper limit of the *m*-sum. As for Case 3, these bounds are not tight for the example considered below, so this case was also studied via simulation. Simulation for all possible *A* and *B* inventories would have required an excessive amount of computer time. Results from Case 4, for the numerical examples considered here, suggested study of a reduced set of possible inventories whose simulation properties should reveal the general properties of this design point of view.

Case 6 is, finally, the simplest kind of redundant spares situation in which partially failed *B*'s are discarded, but good spare *B*'s may be used for either module function. The main emphasis of this

paper has been on the possible advantage of testing to find partially good modules which are usable in the simpler function. This case, without that option, serves as an additional standard for comparison. The adequacy of a set of *KB B* spares and *KA A* spares is simply

$$P(\text{adequacy}) = \sum_{l=0}^{KB} \left[P(NB=l) \sum_{m=0}^{KA+KB-l} P(NA=m) \right]. \tag{9}$$

As in Case 4 (see Equation (8)) the interacting spares structure requires direct comparison of all possible inventories to get undominated solutions.

4. RESULTS

Calculations were carried out for several different sets of cost and reliability parameters. In each case 20 equipments were needed to operate for 10 years. Results are presented here for a single representative case.

Table 3 lists the adequacy and cost results for the "incomplete" optimal design in Case 1 with two module types for separate functions. The results for Case 2 using single module types are also shown there. Only one curve is shown for Case 2 because the upper and lower bounds from (7a) and (7b) coincide for this numerical example. The retrieval capability here means that there will essentially always be an adequate number of spares for partial *B* failure replacements—for either the optimistic or pessimistic failure ordering. The semiconductor design parameters listed there are identified by the symbols (see Appendix A):

ANOS = Number of oxide steps,
ANCC = Number of contact cuts,
TMW = Heat dissipation in milliwatts,
PN = Number of active pins,
COMP = Number of components.

The solid curves in Graph 1 contrast these two spares philosophies by plotting the information from Table 3 with probability of adequacy vs total cost. It is seen that at high adequacy levels, where many spares are needed, the better of these two policies stocks only one kind of spares.

It should be noted that, e.g., Case 1 with $B=1, A=0$ is different from Case 2 with $B=1$ for two reasons. The former is less expensive because it uses *A* modules in *A* functions, and the latter has a higher adequacy probability due to its greater flexibility in using *B* modules for *A* functions.

TABLE 4. *Simulation Results—One Module—Case 3*

Number of *B*	Adequacy level-2	Adequacy level-3	Adequacy correction	Corrected adequacy	Case 2 analytical
1.0	0.264	0.104	−0.020	0.124	0.284
2.0	0.568	0.342	0.028	0.314	0.540
3.0	0.762	0.554	0.008	0.546	0.754
4.0	0.884	0.740	−0.005	0.745	0.889
5.0	0.966	0.872	0.009	0.863	0.957
6.0	0.988	0.964	0.003	0.961	0.985
7.0	0.996	0.988	0.000	0.988	0.996
8.0	1.000	0.998	0.001	0.997	0.999
9.0	1.000	0.998	0.000	0.998	1.000

Table 4 shows simulation results for the single module policy of Case 3, which does not permit retrieval of completely good B modules that have been used in A functions. The analytical adequacy bounds have not been plotted in this case since the simulation results are representative because they lie roughly midway between the respective bounds. This policy gives lower adequacy for the same cost when compared to the more efficient, but less realistic, policy which does allow retrieval, as shown by the dotted curve in Graph 1. Columns 2 and 3 in Table 4 are adequacy probabilities for the "retrieval" and "nonretrieval" policies, as determined by averaging 500 realizations. Comparison of Column 1 of Table 4 with data in Table 3 shows the simulation error for the analyzable case. Using that error as a correction to Column 2 produces the corrected Monte Carlo adequacy estimate for the latter case. Column 6 repeats the adequacy numbers for the first case from Table 3 (Column 4 + Column 6 = Column 2.)

TABLE 5. *Two Kinds of Spares with Retrieval—Case 4*

ANOS	ANCC	TMW	COMP	PN
219.0	103.00	5.60	35.00	10.00
437.0	229.00	8.31	50.00	16.00

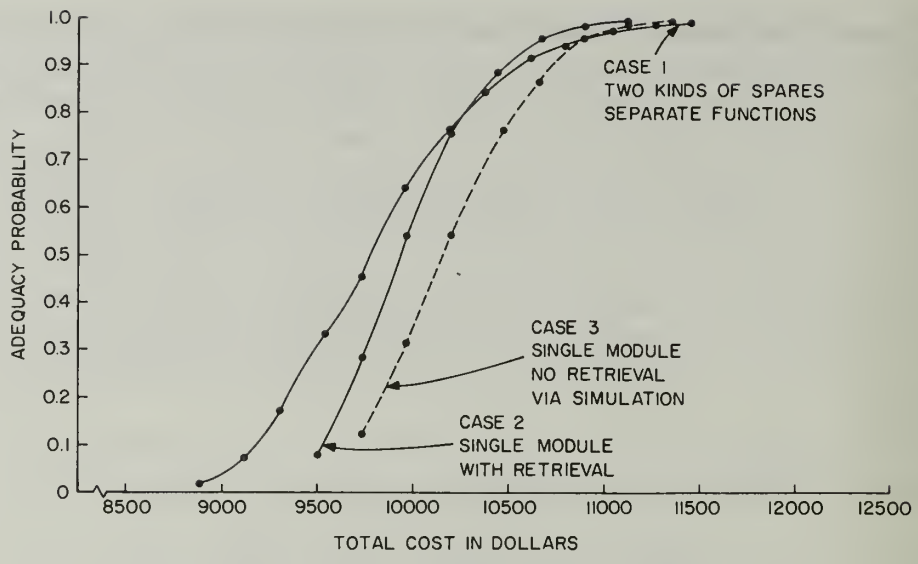
NE = 20.0 ANL = 10.0

Lambda (1) = 0.072

Lambda (2) = 0.148

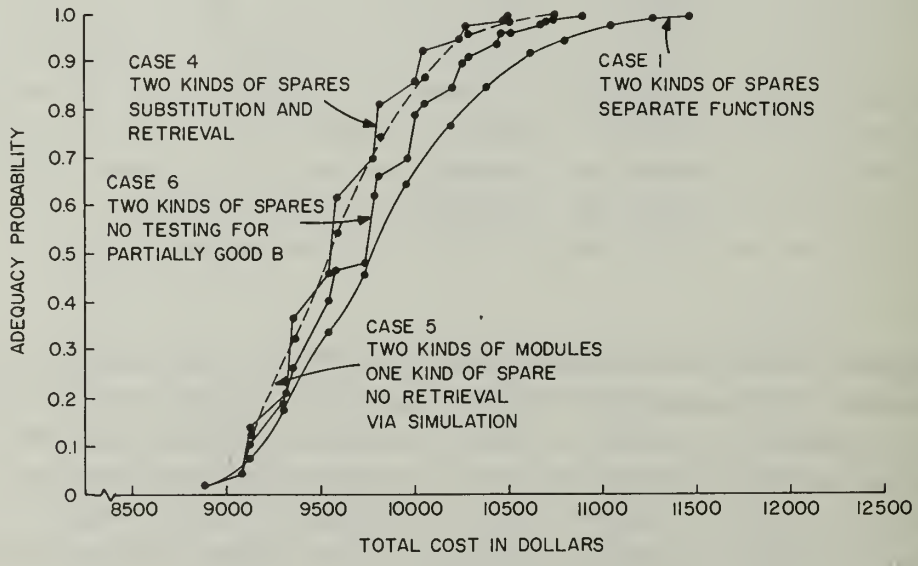
Cost	Reliability	Number. Mod. A	Number Mod. B
0.8886E 04	0.2119E−01	0	0
0.9076E 04	0.4791E−01	1	0
0.9118E 04	0.1384E 00	0	1
0.9308E 04	0.2114E 00	1	1
0.9350E 04	0.3662E 00	0	2
0.9540E 04	0.4598E 00	1	2
0.9582E 04	0.6184E 00	0	3
0.9772E 04	0.6954E 00	1	3
0.9814E 04	0.8101E 00	0	4
0.1000E 05	0.8565E 00	1	4
0.1005E 05	0.9203E 00	0	5
0.1024E 05	0.9423E 00	1	5
0.1028E 05	0.9712E 00	0	6
0.1047E 05	0.9799E 00	1	6
0.1051E 05	0.9909E 00	0	7
0.1070E 05	0.9938E 00	1	7
0.1074E 05	0.9974E 00	0	8

Table 5 shows the undominated policies for Case 4 in which the original equipment uses A modules for the A function and B modules for the B function, and where both kinds may be stocked as spares, but B's may be used for A functions. As in Case 2, these numbers correspond to both the upper and lower bound expressions (8a) and (8b). The solid curves in Graph 2 show that these results are uniformly superior to those for Case 1, as expected. Another interesting property is that while some undominated policies do call for one spare A module, it is invariably true that a slightly more expensive all-B-module inventory has a much better adequacy probability.



GRAPH 1

This last observation motivated restriction of the simulation study of Case 5 to those inventories containing no A modules. In view of the no-interchange discipline, this case is identical to Case 3 except that the *original* equipment is less expensive here because it uses A modules for A functions. Thus the dotted curve in Graph 2 is simply a shifted version of the dotted one in Graph 1. The points on this curve lie roughly midway between the corresponding analytical upper and lower bounds, so those bounds have not been plotted here. This more realistic situation does show somewhat smaller adequacy probabilities than do the Case 4 points corresponding to the same spares inventories.



GRAPH 2

TABLE 6. *Two Kinds of Spares; No Use of Partially Good B Modules—Case 6*

ANOS	ANCC	TMW	COMP	PN
219.0	103.00	5.60	35.00	10.00
437.0	229.00	8.31	50.00	16.00

NE = 20.0 ANL = 10.0
Lambda (1) = 0.072
Lambda (2) = 0.148

Cost	Reliability	Number Mod. *A*	Number Mod. *B*
0.8886E 04	0.2119E−01	0	0
0.9076E 04	0.4791E−01	1	0
0.9118E 04	0.1028E 00	0	1
0.9308E 04	0.1890E 00	1	1
0.9350E 04	0.2602E 00	0	2
0.9540E 04	0.4009E 00	1	2
0.9582E 04	0.4624E 00	0	3
0.9730E 04	0.4781E 00	2	2
0.9772E 04	0.6174E 00	1	3
0.9814E 04	0.6573E 00	0	4
0.9962E 04	0.6965E 00	2	3
0.1000E 05	0.7868E 00	1	4
0.1005E 05	0.8075E 00	0	5
0.1019E 05	0.8482E 00	2	4
0.1024E 05	0.8950E 00	1	5
0.1028E 05	0.9039E 00	0	6
0.1043E 05	0.9335E 00	2	5
0.1047E 05	0.9538E 00	1	6
0.1051E 05	0.9571E 00	0	7
0.1060E 05	0.9741E 00	2	6
0.1070E 05	0.9816E 00	1	7
0.1074E 05	0.9827E 00	0	8
0.1089E 05	0.9909E 00	2	7
0.1093E 05	0.9933E 00	1	8
0.1097E 05	0.9936E 00	0	9

Finally, the results for Case 6, in which partially good *B*'s are discarded, are shown in Table 6 and Graph 2. As in Case 4, while some undominated solutions include one or two *A* modules, slight cost increases result in all *B* inventories with much better adequacies. Comparison of Cases 5 and 6 shows the improvement possible by using partially good *B* modules. (This comparison should be tempered by recalling that the costs for isolating partially and completely failed modules have not been included.)

5. CONCLUSIONS AND FURTHER INVESTIGATIONS

Studies of cost vs adequacy tradeoffs have been carried out for several spares philosophies, some of which take advantage of redundancy between modules needed for different functions. Numerical results for a very simplified system involving only two kinds of module functions display some interesting trends.

The most cost-effective approach uses both kinds of modules, but allows substitution of the more complex function for the simpler one. One might expect that for high adequacy several spares of each

kind would be called for, since several failures in each function might be anticipated. However, the present example calls for essentially stocking only the complex module. A possible interpretation is that high adequacy requires provision for relatively unlikely events such as several failures of only one kind of module function.

It is also interesting to note that a completely single module approach (included in the original equipment) can be better than the conventional approach using two kinds (with each restricted to use in a single function). This is the case for the high adequacy region in Graph 1. It is possible for some module costs and reliabilities, as well as manufacturing setup costs for separate modules, that the single module approach might even be preferable to the redundant two module cases (4, 5 and 6).

It is perhaps worthwhile to list many of the simplifying assumptions which could be relaxed when similar studies are carred out for more realistic situations.

The use of a partially failed module for a simpler function requires availability of test points for isolating the failure as well as equipment and manpower for performing such tests. Inclusion of such test points will increase module cost and module failure rate. The testing operation itself will introduce additional costs. Failures caused by repairmen during testing or replacement might also be included in a more detailed study.

The cost model used here is specifically for physically small integrated circuits in which, e.g., shelf costs do not depend on the number of spare modules, but merely on the number of different bins needed for separate module types. Modification to cost models for other kinds of redundant components should be straightforward.

One simple extension of the above analysis would be to a system where the number m of A modules and the number n of B modules are unequal ($m \neq n$). Another generalization, which could be introduced in the simulation analysis, might assume that some extra cost is incurred for interchange of a partially failed B and a good B module.

The above analysis assumed a constant hazard failure model. The simulation study could also be carried out on the same lines for nonconstant (e.g., linearly decreasing, linearly increasing, or Weibull) hazard models. If the hazard is only slowly varying, the control variate method which employs a constant hazard reference model should still be effective in producing adequacy estimates with low variance.

APPENDIX A—RELIABILITY MODEL

The failure rate prediction procedure from Reference 4 is illustrated in the flow chart shown in the Figure A–1.

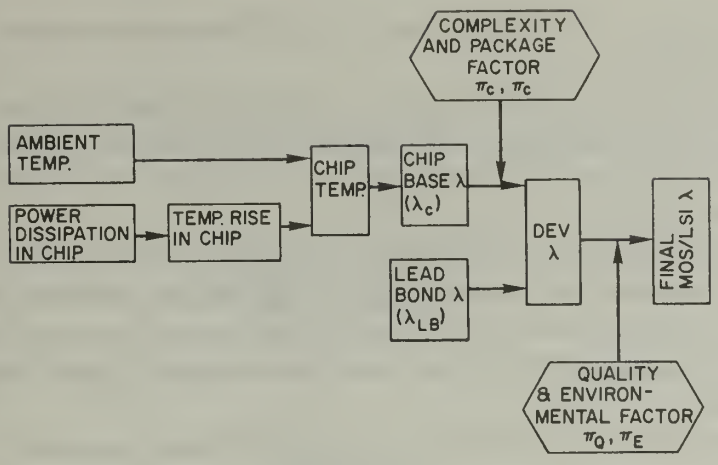
The first four steps (A, B, C, D) compute chip temperature from ambient temperature and rise of temperature due to power dissipation in the chip.

Using Figure A–2 the chip base failure rate (E) is then computed in accordance with the Arrhenius expression $\lambda_c = 0.98e^{-(2298/T)}$ from the RADC notebook procedure.

The package factor (π_p) is assigned a value 2.5 to represent the larger and nonstandard package used for the LSI device.

The complexity factor (π_c) is compiled as follows.

(1) Make a count of total oxide steps and contact cuts in the chip to arrive at a measure of complexity of the circuitry in the chip.



$$\lambda_{MOS/LSI} = \left[(\lambda_C \cdot \pi_C \cdot \pi_P) + \lambda_{LB}\right] \pi_Q \cdot \pi_E$$

FIGURE A–1. Flow chart of MOS/LSI device reliability failure rate prediction procedure.

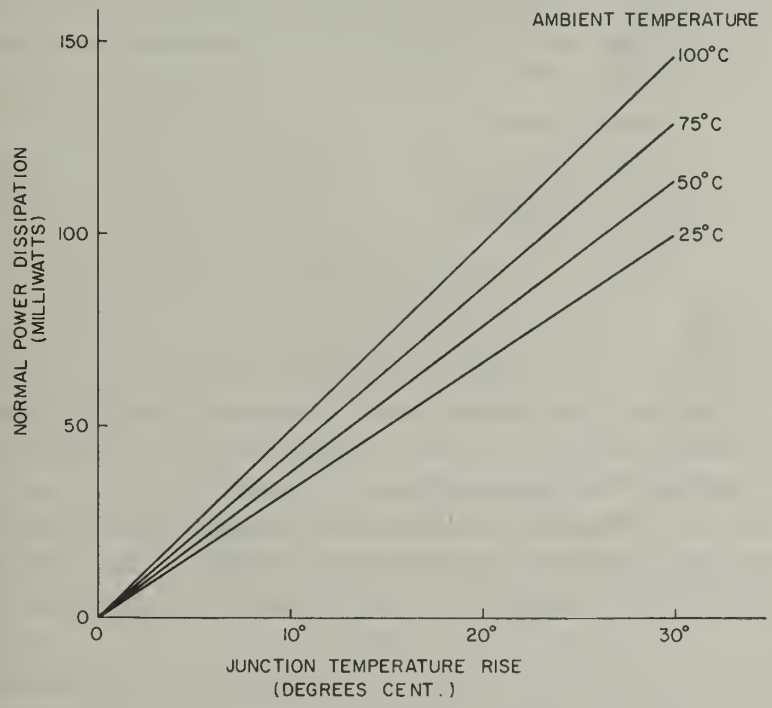


FIG. A-2: – CHIP TEMPERATURE RISE DUE TO POWER DISSIPATION

FIGURE A–2. Chip temperature rise due to power dissipation.

Oxide Step is defined as the MOS gate, consisting of an area of very thin oxide which covers and modulates the MOS channel.

Contact Cut is defined as a vertical cut through the oxide layers for the purpose of making an electrical connection between a diffusion layer and a metallic layer.

(2) For the total count of oxide steps and contact cuts, N, determine the value of complexity factor by dividing by 30; $\pi_c = N/30$.

The base failure rate is then added to introduce the failure probability of lead bond failures. The base failure rate for the lead bonds, λ_{LB}, is determined by multiplying the number of active terminals by 0.00007 percent per thousand hours for aluminum wire ultrasonic bonds, or by 0.00013 for gold wire thermal compression bonds.

The resulting failure rate for a device (H) is then multipled by the factors π_Q and π_E to introduce the effects of quality level and environment resulting in the desired failure rate (K).

A quality factor, π_Q, is assigned a value based on the attention given to quality provisions in wafer processing and device packaging. With proper application of quality controls, a quality factor of 1.0 should be considered.

The environmental factor, π_E, is determined from the RADC procedure in accordance with Table A–1.

TABLE A–1. *Environment Adjustment Factor*

Environment	π_E
Laboratory	1.0
Satellite, orbit	1.5
Ground, fixed	2.0
Ground, portable	5.0
Ground, mobile	7.0
Airborne, inhabited	5.0
Airborne, uninhabited	7.0
Satellite, launch	8.0
Missile	10.0

BIBLIOGRAPHY

[1] Barlow, R. and F. Proschan, *Mathematical Theory of Reliability* (New York, Wiley, 1965).

[2] Caponecchi, A. J. and P. A. Jensen, "A Partitioning Technique for Obtaining Solutions to the Modularization Problem," Nav. Res. Log. Quart. *21*, 13–40 (1974).

[3] Messinger, M. and M. L. Shooman, "Techniques for Optimum Spares Allocation: A Tutorial Review," IEEE Transactions on Reliability, *19*, 156–166 (Nov. 1970).

[4] Nichols, E. D., "MOS/LSI Reliability Prediction," Proceedings of the Annual Symposium on Reliability (1972).

[5] Papoulis, A., *Probability, Random Variables and Stochastic Processes* (McGraw–Hill Book Co., 1965).

[6] Shooman, M. L., *Probabilistic Reliability: An Engineering Approach* (McGraw-Hill Book Co., 1968).

[7] Sinkar, S. and L. Shaw, "Redundant Spares Allocation," Report *EER–104*, EE/EP Dept., Polytechnic Institute of New York (Sept. 1973).

TRANSPORTATION TYPE PROBLEMS WITH QUANTITY DISCOUNTS

V. Balachandran and Avinoam Perry

Graduate School of Management
Northwestern University
Evanston, Illinois

ABSTRACT

It is known to be real that the per unit transportation cost from a specific supply source to a given demand sink is dependent on the quantity shipped, so that there exist finite intervals for quantities where price breaks are offered to customers. Thus, such a quantity discount results in a nonconvex, piecewise linear functional. In this paper, an algorithm is provided to solve this problem. This algorithm, with minor modifications, is shown to encompass the "incremental" quantity discount and the "fixed charge" transportation problems as well. It is based upon a branch-and-bound solution procedure. The branches lead to ordinary transportation problems, the results of which are obtained by utilizing the "cost operator" for one branch and "rim operator" for another branch. Suitable illustrations and extensions are also provided.

INTRODUCTION

In the real world, it is a common practice to offer discounts for the purchase of large quantities d/or for shipment of large volumes of a given commodity [6, 2].

In this paper we analyze quantity discount problems representing the "all unit" [5] and the "incemental quantity" discounted transportation problems. The details of the procedure are explained nsidering the "all unit" quantity discount problem in Section II, whereas the associated algorithm presented in Section III. Certain branch selection procedures and heuristics are provided in Secon IV with a corresponding illustration in Section V. Based on this algorithm, we provide extensions Section VI where the "incremental quantity discounted problem" and the "fixed charge transportaon problems" are addressed specifically.

The analysis will focus on the "all unit" quantity discount problem where the methodology conrning the general approach for solving the piecewise linear programming is developed.

Let $\lambda_{ij}^0 = 0, \lambda_{ij}^1, \lambda_{ij}^2, \ldots, \lambda_{ij}^k, \ldots \lambda_{ij}^r \leq \infty$, where $\lambda_{ij}^{k-1} < \lambda_{ij}^k$ (for $k = 1, 2, \ldots, r$), be such that if a antity x_{ij} is shipped from source i to sink j ($i = 1, 2, \ldots, m$), ($j = 1, 2, \ldots, n$) and $\lambda_{ij}^{k-1} \leq x_{ij} \lambda_{ij}^k$, then the per unit cost of the x_{ij} units is c_{ij}^k, and the total cost associated with shipping x_{ij} units $c_{ij}^k x_{ij}$ where $c_{ij}^k > c_{ij}^{k+1}$. This result is illustrated in Figure 1. This problem may be solved with parable nonconvex programming [11], but solving the transportation problem in this manner has e disadvantage of requiring a large number of constraint equations.

Our approach is similar to the one suggested by Falk and Soland [3] for solving the general nonnvex type math program, but, nevertheless, is more specialized and concerned with the special pe of nonconvex math program, namely the piecewise linear program.

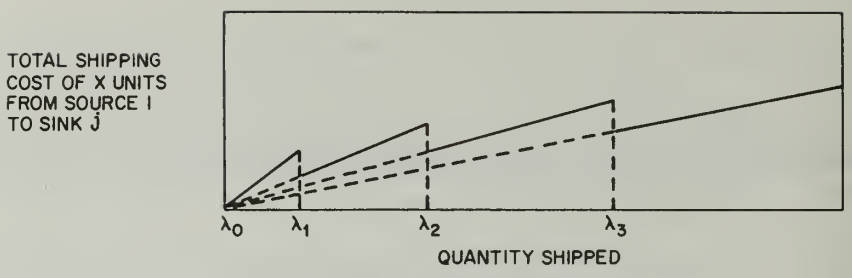


FIGURE 1

II. THE ALL UNIT QUANTITY DISCOUNT TRANSPORTATION TYPE PROBLEM

The transportation type problem with all unit quantity discounts may be formulated as follows:

(1) $$\text{Minimize } Z = \sum_{j\epsilon J} \sum_{i\epsilon I} c^*_{ij}\, x_{ij}$$

Subject to:

(2) $$\sum_{j\epsilon J} x_{ij} = a_i \qquad \text{for } i\epsilon I,$$

(3) $$\sum_{i\epsilon I} x_{ij} = b_j \qquad \text{for } j\epsilon J,$$

(4) $$0 \leqslant x_{ij} \leqslant \lambda^r_{ij} \qquad \text{for } i\epsilon I \text{ and } j\epsilon J;$$

(5) $$c^*_{ij} = \begin{cases} c^1_{ij} & \text{if } 0 = \lambda^0_{ij} \leqslant x_{ij} < \lambda^1_{ij}, \\ c^2_{ij} & \text{if } \lambda^1_{ij} \leqslant x_{ij} < \lambda^2_{ij}, \\ c^K_{ij} & \text{if } \lambda^{k-1}_{ij} \leqslant x_{ij} < \lambda^k_{ij}, \\ c^r_{ij} & \text{if } \lambda^{r-1}_{ij} \leqslant x_{ij} < \lambda^r_{ij} \leqslant \infty. \end{cases}$$

where

$$I = \{1, 2, \ldots, i, \ldots, m\} \text{ set of sources,}$$

(6) $$J = \{1, 2, \ldots, j, \ldots, n\} \text{ set of sinks,}$$

$$R = \{1, 2, \ldots, k, \ldots, r\} \text{ set of cost intervals.}$$

In order to facilitate the presentation, expressions (1)–(5) will be referred to as problem P^*.

The algorithm provided for solving P^* is basically a branch-and-bound type similar to the subtou elimination algorithm of the travelling salesman problem [7]. Here, instead of eliminating infeasibl subtours, we eliminate all infeasibilities due to (5) until complete feasibility is restored.

Let us now define the following "initial Transportation Problem" P_0, which is given b (1)–(4) of P^* (note that constraint set (5) is not included). All c^*_{ij} are replaced with c^r_{ij}. Since c^r_{ij} a the minimum values for every $(i, j)\ \epsilon I \times J$, the solution to problem P_0 is a better than optimal solutio

, in addition, the solution matrix $X = \{x_{ij}\}$ satisfies $\lambda_{ij}^{r-1} \leq x_{ij} < \lambda_{ij}^{r}$ for every i, j, then the solution of oblem P_0 is the optimum solution to problem P^*.

DEFINITION 1. A solution $X = \{x_{ij}\}$ to the relaxed problem (1)–(4) is said to be "interval feasible" all c_{ij}^* used in (1) are implied to be feasible due to the fact that x_{ij} lies in the feasible interval given (5). A solution $X = \{x_{ij}\}$ to (1)–(4) is said to be "interval infeasible" if there is at least one x_{ij} for nich c_{ij}^* does not satisfy constraint set (5).

DEFINITION 2. A "nonterminal solution" is an optimum solution to the relaxed problem (1)–(4) ovided that there exists at least one $\lambda_{ij}^{k-1} \leq x_{ij} < \lambda_{ij}^{k}$ where its $c_{ij}^* = c_{ij}^l$ in (1) and l is any index such that $< l \leq r$.

DEFINITION 3. A "terminal" solution is an interval feasible and optimum solution to the relaxed oblem (1)–(4).

DEFINITION 4. A "better than terminal" solution is defined as a "nonterminal" solution with smaller objective value than that of any terminal solution.

DEFINITION 5. A given solution—"terminal" or "nonterminal"—is said to be "relatively inferior" there exists at least one terminal solution, the objective value of which is smaller.

LEMMA 1. A sufficient condition for $X = \{x_{ij}\}$ to be an optimal solution to problem (1)–(5) is at (a) X is "terminal" and (b) there exists no other "better than terminal" solution. (Proof is obvious.)

The algorithm presented here will always drop all "relatively inferior" solutions and will preserve most one "terminal" solution and some "better than terminal" solutions. If no "better than terinal" solution exists, then the proof to Lemma 1 is immediate.

In addition, the algorithm always possesses a "better than terminal" criterion and proceeds to store "interval feasibility," similar to any dual algorithm which always has a "better than optimal lue" and approaches "primal feasibility."

I. A GENERAL DESCRIPTION OF THE ALGORITHM

In the first stage P_0 is solved.† Thus, if the solution to P_0 is "interval feasible," it is also "terminal" d hence, by Lemma 1, optimal to P^*. If the solution to P_0 is "interval infeasible," then it is "nonrminal". It is also "better than terminal," since there is no other "terminal" solution. The algorithm oceeds as follows: Let x_{ij} be *a value* for which "interval feasibility" is violated. More specifically, ppose x_{ij} is in the interval $\lambda_{ij}^{k-1} \leq x_{ij} < \lambda_{ij}^{k}$ and its associated cost parameter c_{ij}^* is not equal to c_{ij}^k, iis condition leads to two branches (subproblems) as follows:

(i) In branch 1 a lower bound restriction of the form $x_{ij} \geq \lambda_{ij}^{k}$ is *imposed, and* c_{ij}^* remains unchanged.

(ii) In branch 2 the current c_{ij}^* is *replaced by the "interval feasible"* c_{ij}^k, and an upper bound reiction in the form of $x_{ij} < \lambda_{ij}^{k}$ is *implied.*

REMARK 1. If $x_{ij} \geq \lambda_{ij}^{k}$ in both branches at the optimum (i.e. the upper bound in branch 2 is lated), then branch 2 is inferior to branch 1 because c_{ij}^* in branch 1 is *smaller than* c_{ij}^k in branch 2. te that the branching procedure suggested here imposes a lower bound on one branch and a cost nalty rather than an upper bound in the other branch.

The two new transportation problems corresponding to (i) and (ii) are solved. The following sumrizes the different alternate situations:

†We have used the Srinivasan and Thompson [8] algorithm for solving P_0.

(a) Both branches lead to "nonterminal" solutions. Those solutions will be classified as "bette than terminal," and further branching will continue from the one with the lower objective value

(b) Both branches lead to "terminal" solutions, and one branch is "relatively inferior." In this cas the "relatively inferior" solution is dropped, and the other becomes the optimum solution to P^*. (I both solutions have the same objective values, then P^* has alternate optima.)

(c) One branch leads to a "terminal" solution while the other leads to a "nonterminal solution."

(1) If the "nonterminal" solution is "relatively inferior," this solution is dropped. The onl other "terminal" solution is optimal to P^*.

(2) If the "nonterminal" solution is not "relatively inferior," it is classified as "better tha terminal" and further branching continues from it.

In general, at any stage of the branching process there exists at most one "terminal" solution. If in addition, there are some "better than terminal" solutions, then branching continues from that on with the least objective value. All "relatively inferior" solutions ("terminal" or "nonterminal") ar dropped as the process continues. The algorithm terminates if there is no "better than terminal" solution. The one retained "terminal" solution is then optimal to P^*.

Certain discussions of the above procedure are now relevant. First, it is clear that the branchin procedure excludes that noninterval feasibility from any further consideration. Also, we exclude a those "relatively inferior" solutions. Further, corresponding to any index pair (i, j), there is only finite number of intervals. Since there is only a finite number of index pairs, and in each branch a least one such (infeasible) interval is excluded, the algorithm converges in a finite number of steps Secondly, the branching procedure results in a partition of the interval feasible solutions in that subse so many solutions are only implicitly enumerated. Third and most importantly, each subproblem *not completely resolved;* instead, we apply the "Operator Theory" [8] for the transportation proble which is utilized to generate the new solution for each subproblem with minor computational effor

Let x_{st} be one such x_{ij} where the interval feasibility is violated. (In Section IV we provide a heurist for the choice of such x_{st}.) First let us consider (ii) where c^*_{st} is replaced by c^k_{st}. Let $c^k_{st} - c^*_{st} = \delta >$ (due to (5)) be the positive value by which the current cost c^*_{st} is to be increased. The optimal solutio to this problem (where all data of the problem are unchanged except for the new cost $c^* + \delta$) is obtaine by applying the "cell cost operator" [8, 9] δC^+_{st} to the current problem P^*.

Now let us consider the other subproblem (i). Here the only change is the new lower bound impose on x_{st}, i.e., $\lambda^k_{st} \leq x_{st}$. Let $x'_{st} = x_{st} - \lambda^k_{st}$ and $x'_{ij} = x_{ij}$ for all $(i, j) \in [I \times J] - \{(s, t)\}$ such that

$$0 \leq x'_{ij} < \infty \qquad \text{for all } (i, j) \in [I \times J]. \tag{7}$$

Substituting (7) in the current problem we have

$$\text{Minimize} \sum_{j \in J} \sum_{i \in I} c^*_{ij} x'_{ij} + c^*_{st} \lambda^k_{st} \tag{8}$$

$$\text{Subject to} \sum_{j \in J} x'_{ij} = a_i \qquad \text{for } i \in I \text{ and } i \neq s, \tag{9}$$

$$\sum_{j \in J} x'_{sj} = a_s - \lambda^k_{st} \tag{9a}$$

0) $$\sum_{i \in I} x'_{ij} = b_j \qquad \text{for } j \in J \text{ and } j \neq t,$$

0a) $$\sum_{i \in I} x'_{it} = b_t - \lambda^k_{st}$$

1) $$0 \leq x'_{ij} < \infty.$$

The solution to this new problem (8)–(11) can be obtained by utilizing the "Rim Operator Theory"], where a cell rim operator δR^-_{st} is applied by equating $\delta = \lambda^k_{st}$ for the known row s and column t. ote that *instead of resolving* both subproblems, we use the "operator theory of parametric pro- amming for the transportation problem" [8, 9] for computing the solutions to the branch problems.

V. BRANCH SELECTION HEURISTIC AND THE ALGORITHM

Several heuristics may be provided to select an index pair (i, j) for branching from among those at violate the interval feasibility. In this section we provide one that is computationally appealing most cases. This heuristic is based on two stages. In the first stage, we select a node that has the ast objective Z value from among the set S of active "better than terminal" subproblems. Denote this oblem as P_w where $w \in S$ and $Z_w = \min_{u \in S} Z_u$. Once this node selection is made, we then identify that dex pair (s, t) from the set Ω of cells $\{(i, j) \in [I \times J]\}$ where "interval feasibility" is violated. This done by evaluating an "infeasibility index" Q_{ij} for each $(i, j) \in \Omega$ where

2) $$Q_{ij} = (c^k_{ij} - c^*_{ij})\, x_{ij}$$

d determining the most "interval infeasible" index:

3) $$Q_{\max} = Q_{st} = \max_{(i,j) \in \Omega} Q_{ij} = (c^k_{ij} - c^*_{ij})\, x_{ij}.$$

ote that k can be different for different index pairs (i, j). The variable x_{st} associated with $Q_{\max}$ is e variable used for branching in the next stage.

ALGORITHM A1. Algorithm for finding the optimal solution to the "interval feasible" all unit antity discounted nonconvex transportation problem (1)–(5).

Initialization:

STEP 1. Set up the problem P_1 as presented in (1)–(4) with c^*_{ij} in (1) replaced by c^r_{ij}, the smallest st as given in the rth interval $\lambda^{r-1}_{ij} \leq x_{ij} < \lambda^r_{ij} \leq \infty$.

Let Z^* be the objective value of the (current) "terminal" solution. (The initial value assumed by is ∞.) Let $X_1 = \{x^*_{ij}\}$ be the optimal solution to P_1 with basis B_1 and the current optimal cost Z_1. this step we solve the relaxed transportation problem (1)–(4) and obtain the solution with $c^*_{ij} = c^r_{ij}$ for ery (i, j).) Let $S = \{1\}$ denote the set of active problems under consideration, and let $m = 1$ denote e total number of problems generated thus far.

STEP 2. Choose problem P_w for which Z_w is the smallest for $w \in S$. If B_w is interval feasible, i.e. tisfies the constraint set (5) for every $(i, j) \in B_w$, go to Step 8. Otherwise go to Step 3.

STEP 3. Find the set of cells Ω where the cells (i, j) in basis B_w violate interval feasibility (5).

For each $(i, j) \epsilon \Omega$, find the infeasibility index Q_{ij} given by (12), and select the variable to be branched from as in (13). Let the cell corresponding to this variable be (s, t).

STEP 4. Define P_{m+1} as the problem obtained from P_w by increasing the cost of c^*_{st} to c^k_{st}. Problem P_{m+1} and its solution are obtained from problem P_w by applying "cell cost operator" δC^+_{st} (see Srinivasan and Thompson [8]) to P_w where $\delta = (c^k_{st} - c^*_{st}) > 0$. Find the new basis B_{m+1}, and find Z_{m+1} as per δC^+_{st} cost operator application [8, 9].

STEP 5. Define P_{m+2} as the problem obtained from P_w by imposing a lower bounded constraint $x_{st} \geq \lambda^k_{st}$ on the current optimal basis B_w. This solution is obtained by solving the same problem P_w except that the rim conditions (row and column totals) for sth column will be decreased by a value of λ^k_{st}. The solution for P_{m+2} is obtained by applying the cell rim operator δR^-_{st} [8, 9]. The new optimal solution Z_{m+2} = the optimal solution to problem $P_{m+2} + c^*_{st} \lambda^k_{st}$ where c^*_{st} is the same current cost used in problem P_w for the cell (s, t).

STEP 6. Denote the basic optimal solutions to P_{m+1} and P_{m+2} obtained from P_w as X_{m+1} and X_{m+2} with bases B_{m+1} and B_{m+2} respectively. Let Z_{m+1} and Z_{m+2} (suitably modified by another constant) be the corresponding solutions for problems P_{m+1} and P_{m+2}.

STEP 7. Drop w from the set S and include $(m+1)$ and $(m+2)$ in S. Redefine m as $(m+2)$ and go to Step 2.

STEP 8. If $Z_w \geq Z^*$, then P_w is "relatively inferior." Drop w from S, redefine m by $m-1$, and go to Step 2. If $Z_w < Z^*$, then P_w becomes the "terminal" solution, and $Z^* = Z_w$. Compare all existing "better than terminal" solutions in S with the new Z^* to eliminate all new "relatively inferior" problems. If all existing problems in S are eliminated (i.e. $Z_u \geq Z^*$ for every $u \epsilon S$), then $X_w = X^*$ is optimal for (1)–(5). Stop. Otherwise go to Step 2.

It is to be noted that in Step 5, while decreasing the row and column totals by λ^k_{st}, either one of the resultant row or column totals may become negative. This leads to an infeasible subproblem. In such an eventuality that branch is dropped from the set S of active branches.

V. ILLUSTRATION

In this section we will provide a simple illustration of the above algorithm. Consider the following three sources $(m=3)$, four destinations $(n=4)$ transportation problem which has quantity discounted transportation costs. The following Table 1 provides the data of the problem for different c^*_{ij} for the three levels of quantity discounts. We provide an example where cell upper bounds are also imposed.

STEP 1. Table 2 provides optimal solution for the initial problem. In each cell (i, j) the value x_{ij} is written in the northeast corner and U_{ij} in the southwest corner. If a cell is in the basis, then the corresponding c_{ij} cell is circled. Among the nonbasic cells, all those which have x_{ij} at their upper bounds have their corresponding c_{ij} underlined. Those at zero levels are left out without any entry posted in the northwest corner. The optimal dual variables, u_i's for rows and v_j's for columns, are given in the southeast corner. (N denotes a large positive number.) It is easy to check that the solution is optimal for the costs given. Notice that cells (1, 3), (2, 2) and (3, 1) are not *interval feasible.*

STEP 2. Choose P_1; $Z_1 = 945$; B_1 is not interval feasible. Hence go to Step 3.

STEP 3. The set of interval infeasible cells is

$$\Omega = \{(1,3), (2,2), (3,1)\}.$$

feasibility indices are

$$Q_{1,3} = (4-3)\times 25 = 25,$$

$$Q_{2,2} = (6-5)\times 60 = 60 \leftarrow,$$

$$Q_{3,1} = (2-1)\times 25 = 25.$$

hus the variable to branch from is x_{22}.

TABLE 1

Destination / Source	1	2	3	4	Warehouse capacity
1	3 [$20 \leqslant x_{11} < \infty$] 4 [$10 \leqslant x_{11} < 20$] 5 [$0 \leqslant x_{11} < 10$]	6 [$10 \leqslant x_{12} \leqslant 15$] 7 [$5 \leqslant x_{12} < 10$] 8 [$0 \leqslant x_{12} < 5$]	3 [$27 \leqslant x_{13} < 60$] 4 [$15 \leqslant x_{13} < 27$] 5 [$5 \leqslant x_{13} < 15$]	One price bracket 4 Upperbound 30	80
2	Upperbound 20 (one price bracket) 6	5 [$65 \leqslant x_{22} < 75$] 6 [$20 \leqslant x_{22} < 65$] 8 [$0 \leqslant x_{22} < 20$]	8 [$10 \leqslant x_{23} \leqslant 25$] 9 [$5 \leqslant x_{23} < 10$] 10 [$0 \leqslant x_{23} < 5$]	One price bracket 15	90
3	1 [$27 \leqslant x_{31} \leqslant 60$] 2 [$20 \leqslant x_{31} < 27$] 3 [$0 \leqslant x_{31} < 20$]	3 [$60 \leqslant x_{32} < \infty$] 4 [$30 \leqslant x_{32} < 60$] 5 [$0 \leqslant x_{32} < 30$]	10 [$20 \leqslant x_{33} \leqslant 30$] 11 [$10 \leqslant x_{33} < 20$] 12 [$0 \leqslant x_{33} < 10$]	5 [$30 \leqslant x_{34} \leqslant 50$] 6 [$20 \leqslant x_{34} < 30$] 7 [$0 \leqslant x_{34} < 20$]	55
Market demand	70	60	35	60	

TABLE 2

u_i \ v_j	3	0	3	7	Warehouse capacity a_i
0	25 N (3)	6 15	25 60 (3)	30 30 4	80
5	20 20 6	60 75 (5)	10 25 (8)	15 N	90
−2	25 60 (1)	3 N	10 30	30 50 (5)	55
Market demands	70	60	35	60	

X_1 is given above; $Z_1 = 945$; $S = \{1\}$ $m = 1$

STEP 4. P_2 is the problem where cell (2, 2) is contained in the basis and the cost is changed s that it becomes interval feasible.

Problem P_2's solution is obtained by applying cell cost operator [8] δC_{22}^{+} where $\delta = (6-5) = 1$.

Since the current problem is basis preserving, following [8], the new solution becomes $X^{+} = X$ so that x_{ij} values are not altered, and

$$Z^{+} = Z + \delta\, x_{22} = 945 + 1 \times 60 = 1{,}005.$$

Form the set $\Omega = B - \{(2, 2)\} = [(1, 1), (1, 3), (2, 3), (3, 1), (3, 4)]$. Following the notation of [8

$$I_p = I_2 = \{1, 2, 3\};\ I_q = \emptyset,$$

$$J_p = \{1, 3, 4\};\ J_q = \{2\}.$$

The maximum extent μ^{+} to which c_{22} can be increased without changing the basis structure ("basis preserving") will be as in Equation (35) of [8].

$$\mu^{+} = \text{Min} \begin{cases} (c_{ij} - u_i - v_j) & \text{for } (i, j) \epsilon [(I_p \times J_q) \cap LB] \\ (u_i + v_j - c_{ij}) & \text{for } (i, j) \epsilon [(I_q \times J_p) \cap UB]. \end{cases}$$

Now the $(i, j) \epsilon [(I_p \times J_q) \cap LB]$ are cells (1, 2) and (3, 2), and $(i, j) \epsilon [(I_q \times J_p) \cap UB] = \emptyset$.

$$\mu^{+} = \text{Min}\ \{(6 - 0 - 0);\ (3 + 2 - 0)\}$$

$$= 5 \text{ occurring at cell (3, 2).}$$

Thus the basis remains unchanged. The only change occurs in the optimal duals as given below. (Ref to Equation (34) of [8].)

$$u_i^{+} = \begin{cases} u_i + \delta & \text{for } i \epsilon I_p \\ u_i & \text{for } i \epsilon I_q, \end{cases}$$

$$v_j^{+} = \begin{cases} v_j - \delta & \text{for } j \epsilon J_p \\ v_j & \text{for } j \epsilon J_q. \end{cases}$$

Thus $u_1^{+} = 1$; $u_2^{+} = 6$; $u_3^{+} = -1$, and $v_1^{+} = 2$; $v_2^{+} = 0$; $v_3^{+} = 2$ and $v_4^{+} = 6$. The new tableau for P_2 is present in Table 3 with $Z_2 = 1{,}005$.

STEP 5. P_3 is a new branch created by imposing the bound constraint $65 \leq x_{22} < \infty$ on P_1. Th guarantees that (2, 2) will be nonbasic. Now a_2 becomes $80 - 65$ with $b_2 = 60 - 65 < 0$. Since c mand must be nonnegative, this subproblem is infeasible. Thus, this branch and its associated su

anches are deleted from the list of active branches. Figure 2 summarizes the operations up to this int.

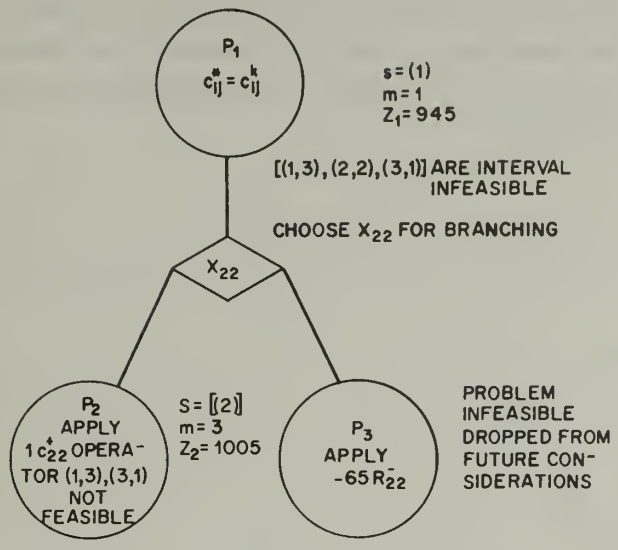


FIGURE 2

STEP 6. P_2, P_3 are given. P_3 is discarded due to infeasibility. X_2 is given in Table 3. $Z_2 = 1{,}005$.

TABLE 3

u_i \ v_j	2	0	2	6	Warehouse capacity a_i
1	25 ③ N	6 15	25 ③ 60	30 4 30	80
6	20 6 20	60 ⑥ 75	10 ⑧ 25	15 N	90
−1	25 ① 60	3 N	10 30	30 ⑤ 50	55
Market demands	70	60	35	60	

STEP 7. (1) is dropped from S and (2) is included in S. m becomes $m+1=2$. Go to Step 2.

STEP 2. P_2 is not interval feasible. Go to Step 3.

STEP 3. $\Omega = \{(1, 3)\ (3, 1)\}$. $Q_{13} = Q_{35} = 25$. Arbitrarily choose x_{13} to branch from.

STEP 4. P_3 is obtained from P_2 by increasing the cost of c_{13} to the interval feasible cost. Following similar cost operation, $1 \cdot c_{13}^{\dagger}$ application, we see the operation is basis preserving. The new cost $Z_3 = 1{,}005 + 1 \times 25 = 1{,}030$. The optimal duals change. The optimal primals do not change. The resultant optimal tableau is given in Table 4.

TABLE 4

v_j / u_i	1	0	2	5	Warehouse capacity a_i
2	25 ③	6	25 ④	30 4	80
	N	15	30	30	
6	20 6	60 ⑥	10 ⑧	15	90
	20	75	25	N	
0	25 ①	3	10	30 ⑤	55
	60	N	30	50	
Market demands	70	60	35	60	

STEP 5. P_4 is obtained from P_2 by imposing the lower bounded constraint $27 \leq x_{13} < \infty$. This creates a new $a_1 = 80 - 27 = 53$ and a new $b_3 = 35 - 27 = 8$. Now we apply cell rim operator δR_1^- with $\delta = 27$. From Theorem 2 of [8], the maximum extent μ^- that this operator can be applied to b basis preserving is 25. But since $\delta = 27$, we follow the method provided by [8]. The new tableau i given in Table 5 with the new optimal primal and dual solutions.

Note that cell (1, 3) has left the basis so that cell (2, 4) is in the basis. The optimal cost is $Z = 930 + 3 \times 27 = 1{,}011$.

Since this problem is not relatively inferior when compared to all pendant branches and, is in terval feasible, it is the terminal solution and hence optimal.

Note that the cell (3, 1) which was interval infeasible earlier became automatically feasible whe the basis change occurred.

The results of branching and bounding are given in Figure 3.

TABLE 5

u_i \ v_j	5	0	2	9	Warehouse capacity a_i
−2	23 ③ N	6 15	3 30	30 4 30	53
6	20 6 20	60 ⑥ 75	8 ⑧ 25	2 ⑮ N	90
−4	27 ① 60	3 N	10 30	28 ⑤ 50	55
Market demands	70	60	8	60	

I. EXTENSIONS

) The Fixed Charge Problem [1]

In this section we outline an algorithm for solving the following fixed charge (transportation) roblem:

5) $$\text{Minimize } Z = \sum_i \sum_j c_{ij}x_{ij} + \sum_k \sum_l f_{kl}y_{kl}$$

Subject to

6) $$\sum_j x_{ij} = a_i \qquad \text{for } i \epsilon I,$$

7) $$\sum_i x_{ij} = b_i \qquad \text{for } j \epsilon J,$$

8) $$y_{kl} = \begin{cases} 0 & \text{if } x_{kl} = 0 \\ 1 & \text{if } x_{kl} \geqslant 1, \end{cases}$$

here $k \epsilon K$, $l \epsilon L$, $i \epsilon I$, $K \subseteq I$, $L \subseteq J$, and $x_{ij} \geqslant 0$.

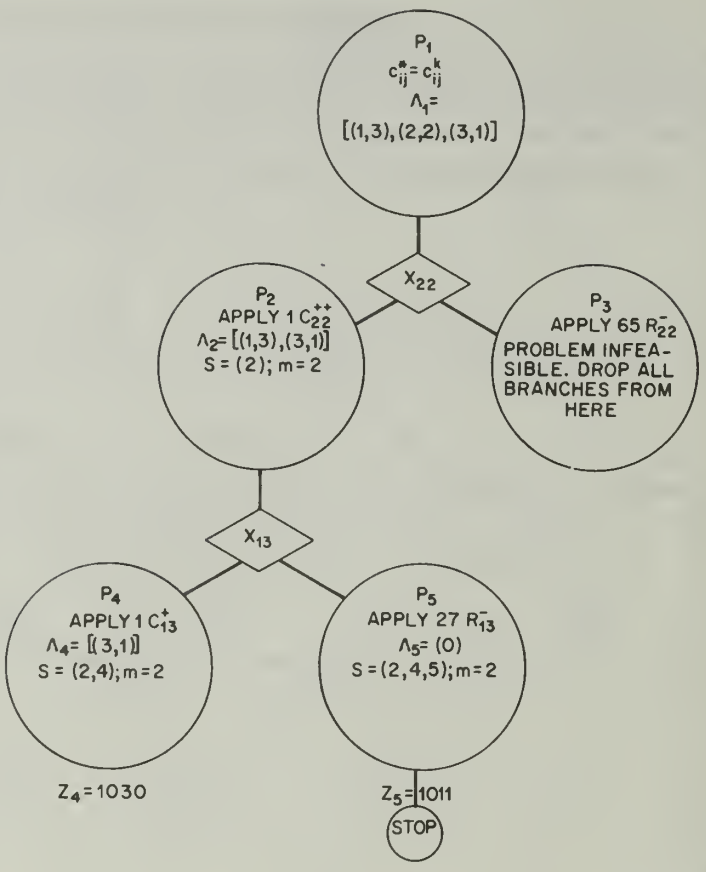


FIGURE 3

The method used for solving the fixed charge problem is identical in principle to the one devise in Section IV for the all unit quantity discount type problem. Here, again, we start by solving a relaxe problem, the solution of which is interval infeasible; then branch-and-bound procedure is applied t retain feasibility. In the following statements the algorithm is summarized:

STEP 1. Let $y_{kl} = 0$ for all $k \epsilon K$ and $l \epsilon L$, and solve the relaxed transportation problem in (15) (18).

STEP 2. If the solution is interval infeasible, i.e. when $x_{kl} > 0$, $y_{kl} = 0$, select the one variab among all interval infeasible variables for which f_{kl} is the largest.

STEP 3. Branch from the variable selected in Step 2 into two new problems: in branch 1 increas c_{kl} to an arbitrary large number making cell (k, l) an inadmissible cell and solve the new problem usi cost operator; in branch 2 introduce a lower bound $x_{kl} \geq 1$ and increase the current value of t objective function by f_{kl}.

STEP 4. Select the one branch with the smallest objective function value among all acti branches.

STEP 5. If the branch selected in Step 4 is interval feasible and terminal, then stop. This is t optimum. Otherwise go to Step 2.

(ii) The Incremental Quantity Discount Problem

The nonconvex cost structure of this problem is shown in Figure 4. To accommodate such

amework, x_{ij}^k must be $\geq \lambda_k$ before x_{ij}^{k+1} (the amount that can be shipped from i to j at a reduced cost $_{j}^{k+1}$) can be positive, etc.

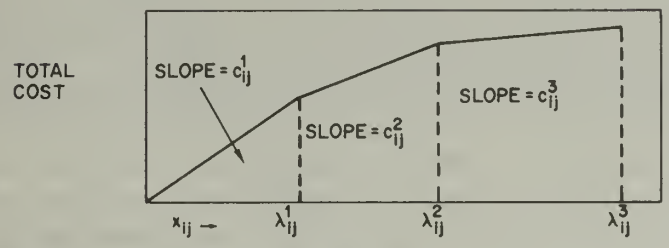


FIGURE 4

The problem can be formulated as follows:

9) $$\text{Minimize} \sum_i \sum_j c_{ij}^k x_{ij} + \sum_i \sum_j f_{ij}^k y_{ij}^k$$

Subject to

0) $$\sum_j x_{ij} = a_i \qquad \text{for } i \epsilon I,$$

1) $$\sum_i x_{ij} = b_j \qquad \text{for } j \epsilon J,$$

2) $$c_{ij}^k = \begin{cases} c_{ij}^1 & \text{if } 0 = \lambda_{ij}^0 \leq x_{ij} < \lambda_{ij}^1 \\ c_{ij}^2 & \text{if } \lambda_{ij}^1 \leq x_{ij} < \lambda_{ij}^2 \\ \cdot & \cdot \quad \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \quad \cdot \\ c_{ij}^r & \text{if } \lambda_{ij}^{r-1} \leq x_{ij} < \lambda_{ij}^r \leq \infty, \end{cases}$$

3) $$y_{ij}^k = \begin{cases} 1 & \text{if } \lambda_{ij}^{k-1} \leq x_{ij} < \lambda_{ij}^k \\ 0 & \text{otherwise} \end{cases}$$

d $x_{ij} \geq 0$ for all $i \epsilon I$ and $j \epsilon J$,

4) $$f_{ij}^k = \left[\sum_{v=1}^{k-1} c_{ij}^v (\lambda_{ij}^v - \lambda_{ij}^{v-1}) \right] - c_{ij}^k \lambda_{ij}^{k-1} \qquad \text{for all } k = 2, 3 \ldots r$$

d $f_{ij}^1 = 0$ for all $i \epsilon I$ and $j \epsilon J$.

A close examination of the above formulation reveals that the incremental quantity discount oblem can be formulated and solved as a generalized fixed charge problem.

omputational Aspects

In the three algorithms, it is shown how one can utilize a modified branch-and-bound procedure

for solving the quantity discounted transportation type problems. These algorithms differ from th one suggested by Falk and Soland [3], whose approach may be characterized by the use of a conve combination of points to approximate the value of c_{ij}^k over a given range; ours is characterized by th use of the marginal cost c_{ij}^k at a given range. However, our approach is more specialized for the no convex but piecewise linear case, while that of Falk and Soland [3] encompasses all nonconvex program

Computationally, very efficient codes for solving transportation problems are available [4, 1 Recently Glover et al. [4] compared different criteria of start procedures, basis changing metho and various algorithms for relative efficiency. A similar study has also been made by Srinivasan a Thompson [10]. It is reported [4] that the procedure of Srinivasan and Thompson is relatively efficie for dense networks, as in the case of the transportation problem. (The average CPU time in Univa 1108 for a 100×100 transportation problem is about 2 seconds.)

In our algorithm, we first solve one transportation problem. The solution of each branch is obtain by utilizing either "cost" or "rim operators" [8, 9]. The relative efficiency of operator theory has been di cussed by Srinivasan and Thompson. Since we are combining relatively efficient procedures reporte elsewhere and, nevertheless, drop all "relatively inferior" branches in the process, there are reasons believe that our algorithms perform well. The number of branches that are generated is dependent up the number of "cost intervals" corresponding to each cell (i, j) and, of course, on the number of row (m) and columns (n). Our numerical illustration should shed some light on the computational a pects. Since this paper can also be viewed as an application of the "Operator Theory" [8, 9], the co putational efficiency of the "Operators" is equally applicable here.

VI. ACKNOWLEDGMENT

The authors express their sincere appreciation for the constructive comments of Professor Srinivasan and a referee, which resulted in a significant improvement in the presentation of the revis version.

REFERENCES

[1] Balinski, M. L., "Fixed Charge Transportation Problems," Nav. Res. Log. Quart. *8*, 41–54 (M 1961).

[2] Central Territory Railroad Tariff Bureau, *E/W–1010, ICC4488* (May 1952).

[3] Falk, J. E. and R. M. Soland, "An Algorithm for Separable Nonconvex Programming Problems Management Science, Vol. *15*, No. *9* (May 1969).

[4] Glover, F., D. Karney, K. Klingman and A. Napier, "A Computation Study on Start Procedur Basis Change Criteria and Solution Algorithms for Transportation Problems," Manageme Science *20*, 793–813 (Jan. 1974).

[5] Hadley, G. and T. M. Whitin, *Analysis of Inventory Systems* (Prentice-Hall, Inc., Englewood Cli N.J., 1963).

[6] Midwest Motor Freight Bureau, Agent, Tariff 350–F, Effective Oct. 14, 1967; Tariff 26–G, Eff tive June 1, 1969; Tariff 40–C, Effective Dec. 13, 1969; and Tariff 1–I, Effective July 27, 19

[7] Shapiro, D., "Algorithms for the Solution of the Optimal Cost Travelling Salesman Probler Sc.D. Thesis, Washington University, St. Louis (1966).

[8] Srinivasan, V. and G. L. Thompson, "An Operator Theory of Parametric Programming for Transportation Problem—I," Nav. Res. Log. Quart. *19*, 205–226 (June 1972).

9] Srinivasan, V. and G. L. Thompson, "An Operator Theory of Parametric Programming for the Transportation Problem—II," Nav. Res. Log. Quart. *19*, 227–252 (June 1972).

9] Srinivasan, V. and G. L. Thompson, "Benefit-Cost Analysis of Coding Techniques for the Primal Transportation Algorithm," Journal of the Association of Computing Machinery (forthcoming).

1] Vogt, L. and J. Even, "Piecewise Linear Programming Solutions of Transportation Costs as Obtained from Rate Traffic," AIIE Transactions, Vol. *4*, No. *2* (June 1972).

AN ASYMPTOTICALLY OPTIMAL INSPECTION POLICY

Dan Anbar

Department of Statistics
Tel-Aviv University
Tel-Aviv, Israel

ABSTRACT

An inspection model in life testing situations is discussed. The system under study is assumed to consist on n independent components all of which fail independently in an exponential fashion. Failures can be discovered only through inspection. The experimenter is assumed to lack the knowledge of the parameter of the exponential distribution. A stochastic sequential inspection policy is suggested which uses the data collected through experimentation to estimate the unknown parameter. It is shown that this policy is asymptotically optimal. Some numerical demonstrations are included.

. INTRODUCTION

There are numerous papers in the literature which deal with the problem of finding optimal spection policies for systems which are subject to failures. A general model is given by Barlow and roschan [3]. Most authors deal with various aspects of replacement and maintenance models. How-ver, with only a few exceptions, one common assumption is that the lifetime distribution is known. To ention some such exceptions, Derman [4] and Roeloffs [5, 6], obtain minimax policies when the fetime distribution is completely unknown, in Derman's case, and partially known, in Roeloffs'.

In this paper we assume that the lifetime distribution is known to be exponential, but with unknown xpected lifetime. We suggest an adaptive sequential inspection policy which is asymptotically op-mal. The procedure goes as follows: A first inspection epoch is determined arbitrarily. At the time the first inspection, the parameter of the (exponential) lifetime distribution is estimated based n the number of failures which occurred during the operation of the system. On the basis of this stimate, the next inspection epoch is determined, a moment at which the estimated parameter is orrected with the aid of the additional data gained, etc. The crucial problem here is the construction the procedure for estimating the expected lifetime. For our purpose the procedure must produce a onsistent sequence of estimated parameter values. This is by no means a trivial task since the usual stimation methods seem to fail to produce procedures with the desired property. The method of stimating the parameter which is used here is a slight modification of a stochastic approximation rocedure due to Albert and Gardner [1] (See also [2]). This procedure yields a sequence of estimators hich is strongly consistent, i.e., converges with probability 1 to the value of the (unknown) parameter. his in turn implies that the sequence of intervals between inspections converges to the optimal in-rval between inspections.

In the last section we describe the results of some computer simulation comparing the performance of the suggested policy with the optimal one.

2. THE MODEL

Consider a system which consists of n, $n \geq 1$, units. The units are working in parallel with identical failure distribution. Let T_i, $i=1, 2, \ldots, n$, denote the lifetime of unit i. The variables $T_i, \ldots, T_n$ are assumed to be independent and identically distributed according to the exponential distribution with (unknown) parameter θ, i.e.,

$$P(T_i \leq t) = F(t) = 1 - e^{-\theta t} \qquad \text{for } t > 0, \qquad \theta > 0. \tag{2.1}$$

We assume that failures can be discovered only through inspection. When inspection takes place, all units are inspected and failing units are replaced by new ones. We assume that a failing unit costs $\$c_0$ per unit of idle time. Replacement of a failing unit costs $\$c_1$ and inspection costs $\$c_2$ per unit. Inspection and replacement are instantaneous.

Since the lifetime is distributed exponentially, inspection and replacement of failing units means renewal of the system. Therefore the optimal inspection policy has the property that the time spans between inspections are constant and independent on the number of inspections which were carried out already. The optimal inspection policy is defined to be the policy which is determined by the time interval τ for which the average cost per unit of time between inspections is minimized.

Suppose we choose an interval of length t between inspections. Denote by $N(t)$ the number of failures which occur during that time. Thus the expected cost per unit of time during that period is given by

$$c(t) = \frac{1}{t} E\left[c_0 \sum_{i=1}^{n} (t - T_i)^+ + c_1 N(t) + nc_2 \right]$$

where

$$(\cdot)^+ = \begin{cases} (\cdot) & \text{if } (\cdot) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$c(t) = \frac{n}{t} \left[c_0 E(t - T)^+ + c_1 F(t) + c_2 \right]$$

where T is a random variable distributed exponentially with parameter θ and $F(t)$ is given by (2.1).

$$E\ (t-T)^+ = \int_0^t (t-u)\, dF(u) = tF(t) - \int_0^t u dF(u)$$

$$= tF(t) - \theta \int_0^t u e^{-\theta u} du = t - \frac{F(t)}{\theta}.$$

Therefore

$$c(t) = \frac{n}{t} \left[c_0 \left(t - \frac{F(t)}{\theta} \right) + c_1 F(t) + c_2 \right]. \tag{2.2}$$

Since $c(t) \to \infty$ as $t \to 0$ and $c(t) \to nc_0$ as $t \to \infty$, there exists a value τ (possibly infinite) for which

(t) is minimized. By differentiating Equation (2.2), we find that a necessary condition for τ to minimize (t) is that it is the solution of the equation

2.3) $$(1+\theta\tau)\, e^{-\theta\tau} = 1 - \frac{c_2}{c_0/\theta - c_1}\,.$$

Let $\phi(x) = (x+1)e^{-x}$, $x \geq 0$. Then $\phi(x)$ is a monotone decreasing function with $\phi(0) = 1$ nd $\phi(x) \to 0$ as $x \to \infty$. Thus a finite solution for (2.3) exists if and only if

$$0 < \frac{c_2}{c_0/\theta - c_1} < 1,$$

r equivalently if and only if

2.4) $$0 < \theta < c_0/(c_1 + c_2)\,.$$

f $\theta > c_0/(c_1+c_2)$, then the optimal policy is to schedule no inspection. We thus assume that (2.4) s satisfied. Thus the optimal policy is determined by

2.5) $$\tau = \theta^{-1}\,\phi^{-1}\left(1 - \frac{c_2}{c_0/\theta - c_1}\right).$$

Since the value of θ is assumed to be unknown, the optimal value τ cannot be computed. We herefore face a statistical problem of estimating θ. In the next section we suggest a method to estimate and, using the estimated values, to construct a sequence of between-inspection times which converge τ with probability 1.

3. THE PROCEDURE

As we have seen in the last section, the optimal procedure is determined by the interval (2.5). Since ϕ is a continuous function, τ is continuous in θ. Thus if $\theta_1, \theta_2, \ldots$, is a sequence of random variables such that $\theta_n \to \theta$ with probability 1, then $\tau_n \to \tau$ with probability 1 where

3.1) $$\tau_n = \theta_n^{-1}\phi^{-1}\left(1 - \frac{c_2}{c_0/\theta_n - c_1}\right), \qquad n = 1, 2, \ldots,$$

provided $\theta_n < c_0/(c_1+c_2)$ with probability 1 for all n. Hence the problem of constructing an asymptotically optimal procedure $\{\tau_n\}$ is equivalent to constructing a strongly consistent sequence of estimators $\{\theta_n\}$. To construct such a sequence we use a stochastic approximation procedure due to Albert and Gardner [1]. We make use of the following convergence result which is proved in Anbar [2].

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $V_1; Y_1, Y_2, \ldots$, be random variables. Denote by $\mathcal{F}_n$ the σ-field generated by $(V_1, Y_1, \ldots, Y_{n-1})$. Let $F_n(x) = F_n(x; w)$, $n = 1, 2, \ldots$, be a random function which is jointly measurable with respect to $(\mathcal{B} \times \mathcal{F}_n)$ where $\mathcal{B}$ is the Borel σ-field of the real line, such that $E(Y_n | \mathcal{F}_n) = F_n(\theta)$ where θ is a fixed real number which is known to belong to an interval (a, b); a and b may be finite or infinite. For $n \geq 1$ define

(3.2) $$V_{n+1}=[V_n-a_n(F_n(V_n)-Y_n)]_a^b$$

where a_n is an $\mathcal{F}_n$-measurable random variable and, for $a<b$,

$$[x]_a^b=\begin{cases} a & \text{if } x\leqslant a \\ x & \text{if } a<x<b \\ b & \text{if } x\geqslant b. \end{cases}$$

THEOREM 1. If the following conditions hold,

(A) for every fixed w, $F_n(x;\,w)$ is a monotone function of the real variable x,

(B) for every $\epsilon>0$ there exists a sequence of nonnegative numbers $\{b_n(\epsilon)\}$ such that for ever $n\geqslant 1$

$$\inf_{\epsilon<|x-\theta|<\epsilon^{-1}} |F_n(x)-F_n(\theta)|\geqslant b_n(\epsilon) \quad \text{with probability 1,}$$

(C) for every $n\geqslant 1$ there exists a nonnegative $\mathcal{F}_n$-measurable random variable A_n such tha with probability 1

$$|F_n(x)-F_n(\theta)|\leqslant A_n|x-\theta| \quad \text{for all } x,$$

(D) for every fixed w

$$\operatorname{sign}(a_n)=\begin{cases} 1 & \text{if } F_n \text{ is nondecreasing} \\ -1 & \text{if } F_n \text{ is nonincreasing,} \end{cases}$$

(E)

(i) $$\sum_n a_n^2A_n^2<\infty$$

(ii) $$\sum_n a_n^2E[(Y_n-F_n(\theta))^2\mid \mathcal{F}_n]<\infty$$

(iii) $$\sum_n |a_n|\,b_n(\epsilon)=\infty \qquad \text{for every } \epsilon>0,$$

with probability 1, then V_n, which is defined by (3.2), converges to θ with probability 1.

To apply Theorem 1 to our situation, we use the following notation. Let τ_1 be a nonnegative randoı variable. Let $N_1, N_2 \ldots$, be a sequence of random variables. Denote by $\mathcal{F}_1$ the σ-field generate by τ_1 and by $\mathcal{F}_k$, $k\geqslant 2$, the σ-field generated by $(\tau_1, N_1, \ldots, N_{k-1})$. Let τ_k, $k\geqslant 1$ be an $\mathcal{F}_k$-mea urable random variable such that $0<A<\tau_n<B<\infty$, where A and B are some fixed number Set $F_k(x)=\max(0,\,1-e^{-x\tau_k})$. Suppose that the conditional distribution of N_k given $\mathcal{F}_k$ is binomia with parameters n and $F_k(\theta)$, where $n\geqslant 1$ is some given integer.

THEOREM 2. Let a_n, $n\geqslant 1$ be a nonnegative $\mathcal{F}_n$-measurable random variable such that

(3.3) (i) $\sum a_n=\infty$ and (ii) $\sum a_n^2<\infty$ with probability 1.

Let θ_1 be an $\mathcal{F}_1$-measurable random variable. For $j \geq 1$ define

.4) $$\theta_{j+1} = [\theta_j - a_j(\theta_1, \ldots, \theta_j)[F_j(\theta_j) - N_j/n]]_a^b.$$

hen $\theta_n \to \theta$ with probability 1.

PROOF. We shall show that the conditions of Theorem 1 are fulfilled.

$$|F_n(x) - F_n(\theta)| = |e^{-\tau_n x} - e^{-\tau_n \theta}| = \tau_n |x - \theta| e^{-\tau_n \theta'}$$

here θ' is some number satisfying $|\theta - \theta'| < |x - \theta|$. Thus, $|F_n(x) - F_n(\theta)| \leq B|x - \theta|$. Therefore condition (C) is satisfied with $A_n \equiv B$, and (E)(i) is implied by (3.3)(ii).

$$\inf_{\epsilon < |x-\theta| < \epsilon^{-1}} |F_n(x) - F_n(\theta)| > e^{-\tau_n \theta} - e^{-\tau_n(\theta+\epsilon)} = e^{-\tau_n \theta}(-e^{-\tau_n \epsilon}) > e^{-B\theta}(1 - e^{-A\epsilon}).$$

hus condition (B) is fulfilled, and (E)(iii) follows from (3.3)(i). Obviously

$$E\left[\left(\frac{N_k}{n} - F_k(\theta)\right) | \mathcal{F}_k\right] \leq \frac{1}{4n},$$

nd condition (E)(ii) follows. This completes the proof.

Theorem 2 provides us with the tool necessary to construct our inspection procedure.

Suppose we know two constants a and b such that $0 < a \leq \theta \leq b < c_0/(c_1 + c_2)$. This guarantees hat Equation (2.5) has a solution. Furthermore, since ϕ is a monotone decreasing function, we also now that the solution τ satisfies $0 < A \leq \tau \leq B < \infty$ where

$$A = b^{-1}\theta^{-1}\left(1 - \frac{c_2}{c_0/a - c_1}\right) \quad \text{and} \quad B = a^{-1}\theta^{-1}\left(1 - \frac{c_2}{c_0/b - c_1}\right).$$

Let θ_1 be an arbitrary random variable taking values in $[a, b]$. Clearly, if an estimated value of is available, θ would be naturally chosen to assume this value. Let

$$\tau_1 = \theta_1^{-1}\phi^{-1}\left(1 - \frac{c_2}{c_0/\theta_1 - c_1}\right).$$

n general, for $k \geq 1$ define

$$\theta_{k+1} = [\theta_k - a_k(\theta_1, \ldots, \theta_k)(F_k(\theta_k) - N(\tau_k)/n)]_a^b$$

nd

$$\tau_{k+1} = \theta_{k+1}^{-1}\phi^{-1}\left(1 - \frac{c_2}{c_0/\theta_{k+1} - c_1}\right).$$

hen by Theorem 2, $\theta_k \to \theta$ and $\tau_k \to \tau$ with probability 1, where τ is the optimal between inspection me.

. SOME NUMERICAL RESULTS

There are certain questions which certainly arise when one wishes to apply our procedure to a

practical situation. The very first question is probably the question of how to choose the sequence $\{a_k\}$. We have no adequate answer to this question. However, some interesting numerical results were obtained with sequences of the form $a_k = Dk^{-1}$. When we restrict ourselves to sequences of that type, the proper choice of D becomes the relevant problem. To judge the performance of the procedure, we stopped iterating as soon as the total sum of the interinspection periods exceeded 365 units of time, i.e., we imagine that we allowed ourselves a period of 1 year of "learning." We call it the learning period. We carried out the simulations to various values of θ, θ_1 and D and the costs c_0, c_1 and c_2. At the end of the learning period we computed the learning cost c_l. That is the difference between the total cost per year using our procedure and the minimal cost per year had we known θ and used the optimal τ from the start. In addition we computed the ratio between the extra cost which we suffer if we stop iterating at the end of the learning period and continue from that time on with the last interinspection period obtained, and the optimal cost. We denote this ratio by R. In the following tables R is given in promills because of the very small values obtained.

All numbers given are averages of 20 repetitions of the sequence of iterations for each set of the parameters. The calculations were carried out on a CDC 6600 at the Computer Unit of Tel Aviv University.

Table 1 summarizes the results of iterations in which the only parameter varied was θ_1. The column $\hat{\theta}$ gives the last estimated value of θ at the end of the learning year.

As we can see, the learning cost is negligible in comparison with the optimal cost, which is 1,063 per day or 387,995 per year. Furthermore, the learning cost is completely insensitive to the starting value θ_1. It starts increasing rapidly only when we approach very closely the endpoints of the interval to which θ is restricted.

TABLE 1. *Dependence of the Process on Starting Value*

$\theta = 0.1$ $(\tau = 2.386)$, $D = 0.5$, $c_0 = 100$, $c_1 = 180$, $c_2 = 20$.
Optimal cost = 1,063 per day.

θ_1	$\hat{\theta}$	c_l (per yr)	R (Promill)	θ_1	$\hat{\theta}$	c_l (per yr)	R (Promill)
0.025	0.1005	366	0.02	0.275	0.1009	69	0.03
0.050	0.1003	108	0.02	0.300	0.1021	71	0.03
0.075	0.1000	59	0.02	0.325	0.1023	93	0.04
0.100	0.0992	50	0.02	0.350	0.1007	65	0.02
0.125	0.1002	50	0.02	0.375	0.1011	52	0.03
0.150	0.1005	58	0.03	0.400	0.1012	96	0.04
0.175	0.1013	98	0.03	0.425	0.1012	134	0.02
0.200	0.1002	72	0.02	0.450	0.1012	217	0.03
0.225	0.1002	80	0.04	0.475	0.1015	589	0.04
0.250	0.1008	71	0.03	0.500	0.1080	25,488	0.17

Table 2 summarizes the results of the following simulation. For various values of θ, we carried out iterations for many different values of D. We then looked for the value of D (D_{opt}), which gives rise to the smallest value of c_l. As we can see, D_{opt} is obtained invariably in the neighborhood of 0.15 independently of the real value of θ.

Table 3 gives the same information as Table 2 with different sets of costs and fewer values of θ.

TABLE 2. *Best Value of* D *as* θ *Varies*

$c_0 = 25, \quad c_1 = 180, \quad c_2 = 20, \quad \theta_1 = 0.05$

θ	$\hat{\theta}$	Opt. Cost (per day)	D_{opt}	c_l (per year)	R (Promill)
0.01	0.0114	153	0.15	139	0.45
0.02	0.0199–0.0202	228	0.10–0.20	76–127	0.14–0.23
0.03	0.0298–0.0304	301	0.10–0.15	23–26	0.03–0.07
0.05	0.0491–0.0503	429	0.10–0.20	2–3	0.00
0.08	0.0793	590	0.15	22	0.03
0.10	0.0966	677	0.15	218	0.40

TABLE 3. *Best Value of* D *as* θ *Varies*

I. $c_0 = 100, \quad c_1 = 180, \quad c_2 = 20, \quad \theta_1 = 0.05$

θ	$\hat{\theta}$	Opt. Cost (per day)	D_{opt}	c_l (per yr)	R (Promill)
(*)0.10	0.1004	1,063	0.50	51	0.02
0.30	0.1004	2,262	0.50	122	0.01
0.40	0.3944	2,709	0.50	153	0.13

II. $c_0 = 50, \quad c_1 = 70, \quad c_2 = 30, \quad \theta_1 = 0.05$

0.20	0.1952–0.2028	980	0.30–0.50	170	0.00
0.40	0.3961	1,394	0.50	59	0.07

((*) This row was computed with $\theta_1 = 0.3$)

In Table 3 we gave two different sets of costs for both of which $c_0/(c_1+c_2) = 0.5$. In both cases we serve that D_{opt} is about 0.50 independently of θ. This may suggest that D_{opt} depends mainly upon e ratio $c_0/(c_1+c_2)$. It is hard to conclude this on the basis of the above data. However, this impression as strengthened by other data which are not reported here.

One other observation which we can make on the basis of the given data is that the learning costs e indeed negligible. This may reflect the situation that the model altogether is not very sensitive to viations from optimality.

CONCLUDING REMARKS

As we have mentioned before, the problem of finding the optimal sequence $\{a_n\}$ is not at all lved. We have no doubt that this is a problem of practical and theoretical interest.

It is also worthwhile to remark that at every stage of the iteration process we make the adjustment the estimated value of θ only on the basis of the data found at the previous stage. We have tried use the intuitive estimator for θ which is based on all the available data, but were unsuccessful in e attempts to prove its convergence. Furthermore, computer simulations did not show any indication at the intuitive estimator does converge. Thus we believe that the straightforward intuitive approach ll fail to produce consistent estimators.

6. ACKNOWLEDGMENT

I wish to thank Professor Shelemyahu Zacks for arousing my interest in the problem and for his constant encouragement. I am also indebted to Mr. Jacob Shaya who faithfully carried out all the computer simulations and made many helpful suggestions regarding the numerical part of this paper.

REFERENCES

[1] Albert, A. E. and L. A. Gardner, "Stochastic Approximation and Nonlinear Regression," Research Monograph No. *42* (The M.I.T. Press, Cambridge, Massachusetts, 1967).

[2] Anbar, D., "An Application of a Theorem of Robbins and Siegmund," Technical Report No. *49*, Department of Statistics, Tel Aviv University, Israel (Aug. 1974). To appear, Ann. of Statistics.

[3] Barlow, R. E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1967).

[4] Derman, C., "On Minimax Surveillance Schedules," Nav. Res. Log. Quart. *8*, 415–419 (1961).

[5] Roeloffs, R., "Minimax Surveillance Schedules With Partial Information." Nav. Res. Log. Quart. *10*, 307–322, (1963).

[6] Roeloffs, R. "Minimax Surveillance Schedules for Replaceable Units." Nav. Res. Log. Quart. *14*, 461–471 (1967).

SELECTION OF THE OPTIMAL SETUP POLICY

Shaul P. Ladany

Ben-Gurion University of the Negev
Beer Sheva, Israel

and

Dina N. Bedi

Baruch College

City University of New York

ABSTRACT

A model is developed taking into consideration all the costs (namely cost of sampling, cost of not detecting a change in the process, cost of a false indication of change, and the cost of readjusting detected changes) incurred when a production process, using an unscheduled setup policy, utilizes fraction-defective control charts to control current production. The model is based on the concept of the expected time between detection of changes calling for setups. It is shown that the combination of unscheduled setups and control charts can be utilized in an optimal way if those combinations of sample size, sampling interval, and extent of control limits from process average are used that provide the minimum expected total cost per unit of time.

The costs of a production process that uses unscheduled setups in conjunction with the appropriate optimal control charts are compared to the costs of a production process that uses scheduled setups at optimum intervals in conjunction with its appropriate control charts.

This comparison indicates the criteria for selecting production processes with scheduled setups using optimal setup intervals over unscheduled setups. Suggestions are made to evaluate the optimal process setup strategy and the accompanying optimal decision parameters, for any specific cost data, by use of computer enumeration. A numerical example for assumed cost and process data is provided.

NTRODUCTION

Fraction-defective control charts can be used for quality control in many production processes. typical example is their application in the manufacture of pipe fittings by pressure casting, where the action of leaky castings can be considerable due to surface porosity. In order to select the most dvantageous control procedure using a fraction-defective control chart, management has to decide in dvance the setup policy applied to the manufacturing process. In the present context "Setup" means e adjustment of the production process, usually by screws, levers, etc., which it is customary to perrm at the start of production following any type of idle period (i.e., lunch break). The alternatives vailable are (a) scheduled setups at fixed intervals in addition to unscheduled setups performed beveen the scheduled setups whenever a change of a certain magnitude in the process average is

detected, and (b) unscheduled setups whenever a change of a certain magnitude in the process averag is detected.

The selection of the setup policy is complicated by the fact that there is a mutual dependenc between the setup policy and the proper control procedure using a fraction-defective control chart Each setup policy requires a different optimal control chart procedure that will minimize the total cos incurred to control current production. Thus the selection of setups should be done by comparing th total costs per unit of time obtained when optimal control chart procedures, corresponding to the differ ent setup policies, are utilized. Recently Ladany [8] has developed a model which determines the opti mal decision parameters required to assure the minimal total expected cost when fraction-defective control charts are used to control a process with *scheduled* setups at fixed intervals. In the following, a model will be initially developed to derive the optimal parameters for minimum total cost per unit o time for processes with *unscheduled* setups that use fraction-defective control charts, and then a approach will be outlined based on the total cost per unit of time for selection of the best combinatio of setup policy and control chart procedure.

The optimal economic design of quality control procedures has been investigated by variou authors. Duncan [2, 3], Gibra [4], Knappenberger and Grandage [7], and Baker [1] have develope models for the $\bar{X}$ chart; Goel et al. [5] have developed an algorithm for computing the optimal param eters for Duncan's model; Montgomery and Klatt [9] have dealt with multivariate extension of the $\bar{X}$ chart; while Hall and Eilon [6] have considered optimal resetting policies implying the use of the $\bar{X}$ chart. On the other hand, the economic design of p charts got only the attention of Ladany [8].

The present approach differs from the one adopted in the previous work. It is based on the cor cept of the expected time between unscheduled setups, as opposed to the average number of stoppage in the fixed time interval between two scheduled setups. The present approach and Duncan's [classical approach are dealing with a single assignable cause. However, they differ drastically not onl because this approach investigates a p versus an $\bar{X}$ chart, but also because, while both intend to ca culate the expected time from setup to detection, the present paper takes into account that when a fals alarm occurs, it is either indistinguishable from a real shift and therefore a setup is performed, or is soon realized that the process has not shifted, which is equivalent to a setup.

The model is based on the assumptions that (a) at the start of the production, the machine produ ing the output to be controlled is setup correctly, (b) there is a given probability that the process wi go out of control during a given *fixed* time interval of any length, (c) the events of going out of contr during the intervals between two consecutive samples are independent, (d) the process will go out control suddenly, changing the mean process fraction-defective by at least a given amount, and (e) detected shift in the process average is corrected immediately and properly.

There are many processes to which the above assumptions apply. A typical example is the pressu casting of pipe fittings. Many chance variables affect the surface porosity of the product and henc determine the average fraction of defectives. Among these chance variables is the pressure, which reduced below a certain limit will suddenly increase the fraction of defectives. This happens whe the hole through which the liquid metal is poured into the mold is suddenly enlarged due to a breakdow caused by metal fatigue. The magnitude of such sudden changes is known from experience or can b calculated from theoretical considerations. The probability of such changes appearing in a given tim period can be measured from historical information. The time period to which this probability is relate might be of any length, but customarily it is the length of a shift, a day, or a week. Pressure casting

ually a continuous process which has no distinguishable end points suited for low cost scheduled ups.

On the other hand, scheduled setups at fixed time intervals disrupting the continuity of the casting uld probably have the same unit cost per setup as unscheduled setups. Therefore such production ocesses are seemingly equally well suited to be setup at any time when changes requiring setup are ected or at the end of scheduled time intervals of equal length. The decision criterion for the selec- n of the preferred setup policy would be the minimum expected cost per unit of time.

IE COST MODEL FOR UNSCHEDULED SETUPS

The model is based on the concept of the total expected cost per unit of time obtained from the al expected cost between two consecutive unscheduled setups divided by the expected length of time ween them. The total expected cost between two consecutive unscheduled setups, TC, consists of r different components:

$$TC = A + B + C + D$$

ere

A = Expected cost of sampling during the time between two setups.

B = Expected cost of a higher rate of defective products for the duration between two setups when the process suddenly goes out of control and this is not detected.

C = Expected cost for the duration between two setups of *false* indications that the process went out of control.

D = Cost of one setup, i.e., readjusting a detected shift.

Since the production process will be controlled by use of a fraction-defective control chart, the lowing three parameters—dependent on the production facilities or decided by management as ue judgment—have to be known:

(a) the mean process fraction-defective, $\bar{p}$

(b) the sudden change in the process fraction-defective which is desired to be detected, d.

(c) the probability that the sudden change in the process fraction-defective, which has to be detected, will occur in the time interval of fixed length, $T - P_s$.

e decision variables encountered in this problem are the following:

(a) the equal length time interval between two consecutive samples, t

(b) size of the sample, n

(c) the extent of control limit from the process average expressed in terms of standard deviations, K.

The fraction-defective control chart will have a center line at $\bar{p}$ and an upper control limit, UCL, at

$$UCL = \bar{p} + K\sqrt{\bar{p}(1-\bar{p})/n}.$$

e lower control limit will not cause stoppages in the process to check for assignable causes when provement in fraction-defectives is indicated, and therefore it has not been considered.

The probability of a type I error, where α = the probability of getting a *false* indication at any mple that the process fraction-defective has changed, exists when the process fraction-defective has

not changed, but the sample fraction-defective, due to chance causes, falls outside the control lim (on the upper side):

$$\alpha = \sum_{x \geq n(UCL)}^{n} \binom{n}{x} \bar{p}^x (1-\bar{p})^{n-x} = P\left(Z \geq \frac{UCL - \bar{p}}{\sigma_{\bar{p}}} \right) \tag{3}$$

where Z is a standard binomial deviate (i.e., $(x - E(x))/(v(x))^{0.5}$, where $x \sim B(n, p)$), since the fra tion-defectives are distributed binomially.

In a similar way, the probability of a type II error, where P_a = the probability of not detecting change of a positive d in the process fraction-defective at the first sample after the change has occurre exists when the change occurs and the fraction-defective of the first sample after the change falls belo the upper control limit:

$$P_a = \sum_{x=0}^{x \leq n(UCL)} \binom{n}{x} (\bar{p}+d)^x (1-\bar{p}-d)^{n-x} = P\left(Z \leq \frac{UCL - (\bar{p}+d)}{\sigma_{(\bar{p}+d)}} \right) \tag{4}$$

$$P_a = P\left(Z \leq \frac{K\sqrt{\bar{p}(1-\bar{p})} - d\sqrt{n}}{\sqrt{(\bar{p}+d)(1-\bar{p}-d)}} \right) \tag{5}$$

where $\sigma_{(\bar{p}+d)}$ is a subscripted standard deviation representing the standard deviation of the chang process.

The probability of the sudden change in the process fraction-defective which has to be detect P_s, is related to a fixed time interval of length T. T might be of any length, customarily the length o shift, a day, a week, etc., but it has to be an interval of time which allows for the collection of unc torted historical data about P_s. It is assumed that the time interval of length T is composed of T s intervals of equal lengths of a unit of time having equal and independent probabilities of getting out control, P_T.

Hence the probability that the sudden change in the process fraction-defective will *not* occur dur the length of time T, $1-P_s$, is equal to the probability that the shift will *not* occur in either of the unit-time subintervals, i.e., to the product of T equal probabilities that the shift will not occur dur a unit-time subinterval, given that at the start of the unit-time subinterval the shift has not yet occurr

$$1 - P_s = (1 - P_T)^T \tag{6}$$

Using the above, it can be shown that the probability that the shift will occur during the length time t between two consecutive samples given that at the start of the interval the shift has not occurred, P_t, is such that

$$1 - P_t = (1 - P_T)^t, \tag{7}$$

and substituting Equation (7) into (6)

$$1-P_t=(1-P_s)^{t/T}$$

$$P_t=1-(1-P_s)^{t/T}.$$

In the model all the costs will be related to the time interval between two consecutive unscheduled ıps. During this period an unknown number of samples will be taken, the samples being spaced equal time intervals t. The schematic descritpion of this sampling procedure is given in Figure 1, wing the difference between the meaning of the consecutive numbering of the samples after the ıp and of the sampling subintervals between samples after the setup.

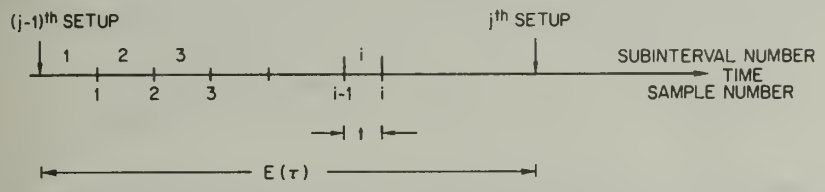


FIGURE 1. Schematic description of the samples and the sampling intervals.

Let's define $G(i)$ as the probability that no shift will occur from the last setup until the start of the sampling interval since the last setup. Since the probability that no shift will occur in the first erval is $1-P_t$, and since the sampling intervals have independent probabilities,

$$G(i)=(1-P_t)^{i-1}.$$

The time between two consecutive setups will be prolonged up to the time that a shift will be ected in the last sample. However, when a false alarm occurs, it is either undistinguishable from a l shift and therefore a setup is performed, or it is soon realized that the process has not shifted l this realization may be equivalent to a setup. Thus a detection of a real shift or of a false alarm become the terminal point which defines the length of a cycle. Therefore, hereafter the union of the mutually exclusive events "real shift" or "false alarm" will just be termed *shift*. In order to cal- ate the expected number of samples during a cycle (between two setups, two false alarms, or any nbinations of them), it is necessary to calculate the expected number of samples after a setup or a ısi setup until the first detection of a shift in a sample, $E(i)$. However, to calculate $E(i)$ it is first essary to define the following:

(a) $D(i)$ = the probability that no shift will be detected on the ith sample,

(b) $H(i)$ = the probability that a shift will be detected on the ith sample.

Figure 2 is a tree diagram describing all the alternative resulting events after the ith sample and ways in which they are obtained, given that the $(i-1)$th sample failed to detect any shift. The set of nts composing the compound event "no shift is detected on the ith sample" are described in Figure 2 events (2), (4), and (6). The set of events composing the compound event "shift is detected on the ith ıple" consist of events (1), (3), and (5).

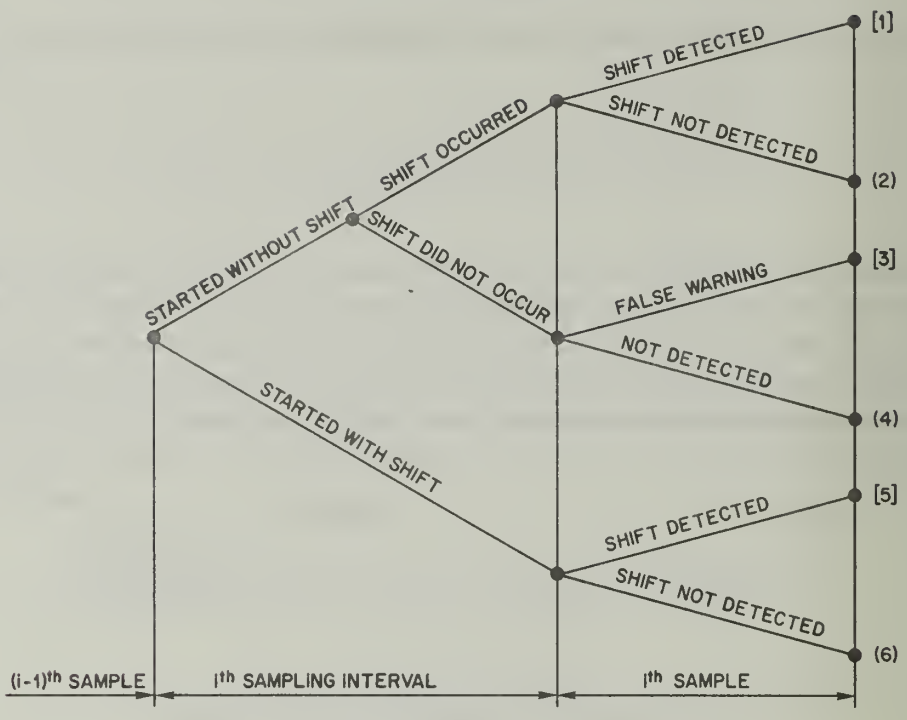


FIGURE 2. Tree diagram of the events after the ith sample.

The union of the mutually exclusive events (2), (4), and (6) is the set of the compound event "
shift is detected on the ith sample." The probability of this event is conditioned by the nondetecti
of a shift during the previous $i-1$ samples since the last setup. As a consequence

$$D(i)=D(i-1)[P((2)\cup(4)\cup(6))]. \tag{11}$$

After substituting the proper probabilities for each of the component mutually exclusive events,
obtain:

$$D(i)=D(i-1)[G(i)P_tP_a+G(i)(1-P_t)(1-\alpha)+(1-G(i))P_a]. \tag{12}$$

This recursion formula can be solved using the assumption for the initial condition that the previ
setup was correct, i.e.,

$$D(0)=1. \tag{13}$$

The union of the mutually exclusive events (1), (3), and (5) is the set of the compound event "sl
is detected on the ith sample." The probability of this event is also conditioned by the nondetect
of a shift during the previous $i-1$ samples since the last setup, providing

$$H(i)=D(i-1)P((1)\cup(5)\cup(3)) \tag{14}$$

and after substituting the appropriate probabilities

$$H(i)=D(i-1)[G(i)P_t(1-P_a)+(1-G(i))(1-P_a)+G(i)(1-P_t)\alpha]. \tag{15}$$

Hence, the expected number of samples after a setup until the detection of a shift $E(i)$ in the ith mple, i.e., during a cycle, will be

$$E(i)=\sum_{i=1}^{\infty} i\,H(i). \tag{16}$$

From Equations (12) and (15) it is evident that $D(i)/D(i-1)+H(i)/D(i-1)=1$ for every $i\geq 1$ d also that $D(i)+\sum_{j=1}^{i} H(j)=1$ for every $i\geq 1$. However, it is assumed that $D(\infty)=0$.

ıe Cost of Sampling

The expected cost of sampling during the cycle, A, equals the product of the expected number of mples during the cycle, the sample size n, and the cost of sampling one item, C_1.

$$A=C_1\,n\,E(i). \tag{17}$$

ıe Cost of Not Detecting Shifts

Let C_2=the cost of not detecting a specified change in the process for the length of the time erval T (to which we have related P_s). The corresponding cost for the length of time t would be where

$$C_2'=C_2t/T. \tag{18}$$

The expected cost of not detecting a specified change in the process which occurred during the st subinterval, assuming that it occurred in the middle of the first subinterval, is

$$G(1)\,P_t\,\frac{C_2'}{2}+G(1)P_tC_2'D(1)+G(1)P_tC_2'D(2)+\ldots\ldots$$

However, the expected cost of not detecting a specified change in the process for the length of ıe of the cycle, the change assumed to occur at the middle of some of the subintervals, B, is

$$B=\sum_{i=1}^{\infty}\frac{G(i)P_tC_2t}{2T}\,[1+2\sum_{j=1}^{\infty}D(j)] \tag{19}$$

ıere the value of C_2' has already been substituted by Equation (18).

ıe Cost of False Indications

During the cycle 0 or 1, false indications might occur. The false indication occurs only when the cle was terminated on the ith sample due to a false indication. Thus, the fraction of false indications

in cycles consisting of i samples, S, equals the conditional probability of obtaining a false indicatio on the ith sample, given that a shift (real shift or false indication) was detected on the ith sample:

$$S=P([3]|[1]\cup[5]\cup[3])=\frac{P[3]}{P([1]\cup[5]\cup[3])}=\frac{D(i-1)G(i)(1-P_t)\alpha}{H(i)}.$$

The expected fraction of false indications in a cycle, $E(S)$, equals the sum of the products of S and the probability that the shift will be detected on the ith sample:

$$E(S)=\sum_{i=1}^{\infty} SH(i)=\alpha(1-P_t)\sum_{i=1}^{\infty} D(i-1)G(i). \tag{20}$$

Let C_3 be the average cost of each false indication. The expected cost of false indications durin a cycle, C, will therefore become

$$C=C_3\alpha(1-P_t)\sum_{i=1}^{\infty} D(i-1)G(i). \tag{21}$$

The Cost of Setups

Let C_4 be the cost of setting up each detected shift when the setups are unplanned and are pe formed at unscheduled intervals. In each cycle there are 0 or 1 setups and the expected fraction o setups is $1-E(S)$. Thus, the expected cost of setups during a cycle is $C_4[1-E(S)]$.

The Total Cost per Unit of Time

The expected total cost per unit of time, when using unscheduled setups, CU, is approximated b total expected cost incurred during the cycle, TC, divided by the expected length of time of a cycl $E(\tau)$. It is clear that even if the total cost random variable is independent of τ, the ratio of their e pected values would not be the expected value of their ratios. Due to the complicated dependencie between the total cost and the "length of time of a cycle" random variables, it seems impractical t derive an analytic expression for the expected value of their ratios. However, the use of the ratio expected values, though not strictly correct, is considered an excellent approximation. The expecte length of time of the cycle equals the product of the expected number of samples during the cycle an the length of time between two samples:

$$E(\tau)=tE(i).$$

Thus

$$CU=\frac{C_1nE(i)+\frac{C_2t}{2T}P_t\sum_{i=1}^{\infty} G(i)[1+2\sum_{j=i}^{\infty} D(j)]+C_3E(S)+C_4[1-E(S)]}{tE(i)} \tag{22}$$

The problem which has to be solved is to find the values of t, n, and K which minimize C $\mathrm{Min}(CU_{n^*,K^*,t^*})$, subject to the conditions of Equations (3), (5), (9), (10), (12), (13), (15), (16), and (2

ıs to be noted that in cases where a false indication is indistinguishable from a real shift and there- a setup is performed, C_3 is C_4, and the sum of the last two terms in the numerator of Equation (22) duced to C_4.

E COST MODEL FOR SCHEDULED SETUPS

In a recent paper [8] a model has been developed to express the total expected cost incurred when ion-defective control charts are utilized to control current production. The basic assumption of the el was that the scheduled setups of the machine producing the output to be controlled are per- ed at *fixed* time periods of a given length T. Accompanying this assumption was the probability btained from historical data, that the process will go out of control during this fixed time period h consists of subintervals in which the probabilities of getting out of control were independent. Using that model, optimal conditions were determined for the combination of sample size, n, uency of sampling, f, and extent of control limits from process average, K, in such a way that mini- total expected cost, $TC_{n,K,f}$, should have been attained for the length of the fixed time period T een two consecutive scheduled setups.

The problem to be solved was as follows:

$$\text{Min } (TC_{n,K,f}) = C_1 nf + \frac{C_2 P_f}{2(f+1)} \sum_{i=0}^{f} F(i) \left[1 + 2P_a \left(\frac{1-P_a^{f-i}}{1-P_a} \right) \right]$$

$$+ C_3 \alpha (1-P_f) \sum_{i=0}^{f-1} F(i) + C_4 P_f \sum_{i=0}^{f-1} F(i)(1-P_a^{f-i}),$$

e α and P_a had to fulfill Equations (3) and (5), while

$$P_f = 1-(1-P_s)^{1/f+1}$$

$$F(i) = \frac{P_a^{i+1} P_f (1-P_f)^i}{1-P_a(1-P_f)} + \frac{1-P_a}{1-P_a(1-P_f)} .$$

However the fixed time interval T between two consecutive scheduled setups (to which the proba- y of getting out of control was related) is not the only possible scheduled setup policy. Any other interval T_r can be considered for the scheduled setup policy, even though it might be possible the cost of performing each scheduled setup after an interval T, which coincides with the end of a , a day, a week, etc., C_5, might be lower than the cost per scheduled setup after an interval T_r which ıpts the production process, C_{5r}.

In order to calculate the minimal total cost for time periods of length T_r, Min $(TC_{n,K,f} | T_r)$, using ation (23), it is necessary to obtain the probability P_{sr} that the process will go out of control during scheduled setup interval T_r. This value of P_{sr}, corresponding to the setup interval T_r, has to be used

instead of P_s in Equation (24). Assuming that each subinterval of any length of time has independe probabilities of getting out of control, according to Equation (9) it is possible to show that

$$P_{sr} = 1 - (1 - P_s)^{T_r/T}. \tag{26}$$

In addition to the replacement of P_s with P_{sr}, it is also necessary to replace C_2 in Equation (23) wi $C_2'' =$ the cost corresponding to the time interval T_r, so that $C_2'' = C_2 T_r / T$.

Thus the total expected cost per unit of time when using scheduled setups at intervals of T_r, C is approximated, as previously for CU, by the total cost incurred between the setups plus the cost the setup divided by the length of the setup interval:

$$CS = \frac{(TC_{n,f,K}|T_r) + C_{5r}}{T_r} \tag{27}$$

The problem to be solved is to find n^*, f^*, K^*, and T_r^* which minimize CS, so that

$$\text{Min}\,(CS_{f,n,K,T_r}) = \frac{\text{Min}\,[(TC_{n,f,K}|T_r) + C_{5r}]}{T_r}. \tag{28}$$

This $\text{Min}\,(CS_{f^*,n^*,K^*,T_r^*})$ has to be compared with $\text{Min}\,(CU_{t^*,n^*,K^*})$ in order to find which set policy has a lower cost per unit of time for a particular set of input parameters. This problem of findi the preferable setup policy and the corresponding optimal decision variables can be solved relativ easily by using the computer screening method outlined in the flow chart of Figure 3. The solution easy since n and f are integers and since t and T_r, though continuous, can be considered in integer ste and all have only a moderately limited practical range. The above considerations also apply to the sel tion of K which can have any positive value, but it is practical to consider only the range from $K=$ to $K=4.0$ in steps of 0.1 or 0.5.

Using this approach, the optimal policy of using unscheduled setups, as well as the optimal poli of using scheduled setups at optimal setup intervals are obtained, providing the corresponding optin decision parameters for each case, n, K, t, or f, as well as the resulting total expected cost per unit time. The setup policy with the lowest cost per unit of time will be selected as the recommended o for any particular set of input data. However, results might also be obtained for "optimal" schedul setup policies at nonoptimal setup intervals, for cases in which nonquantifiable factors might overr insignificant cost advantages.

NUMERICAL EXAMPLE

A manufacturing process had a mean process fraction defective, $p=0.08$; a sudden change $d=0.04$ in fraction defective was aimed to be detected; the length of a shift was $T=8$ hours; it v found that during 60 percent of 8 hour shifts, at least one shift in the adjustment occurred which cal for a setup and thus $P_s=0.6$; and the average time to detect a false warning was 5.2 minutes.

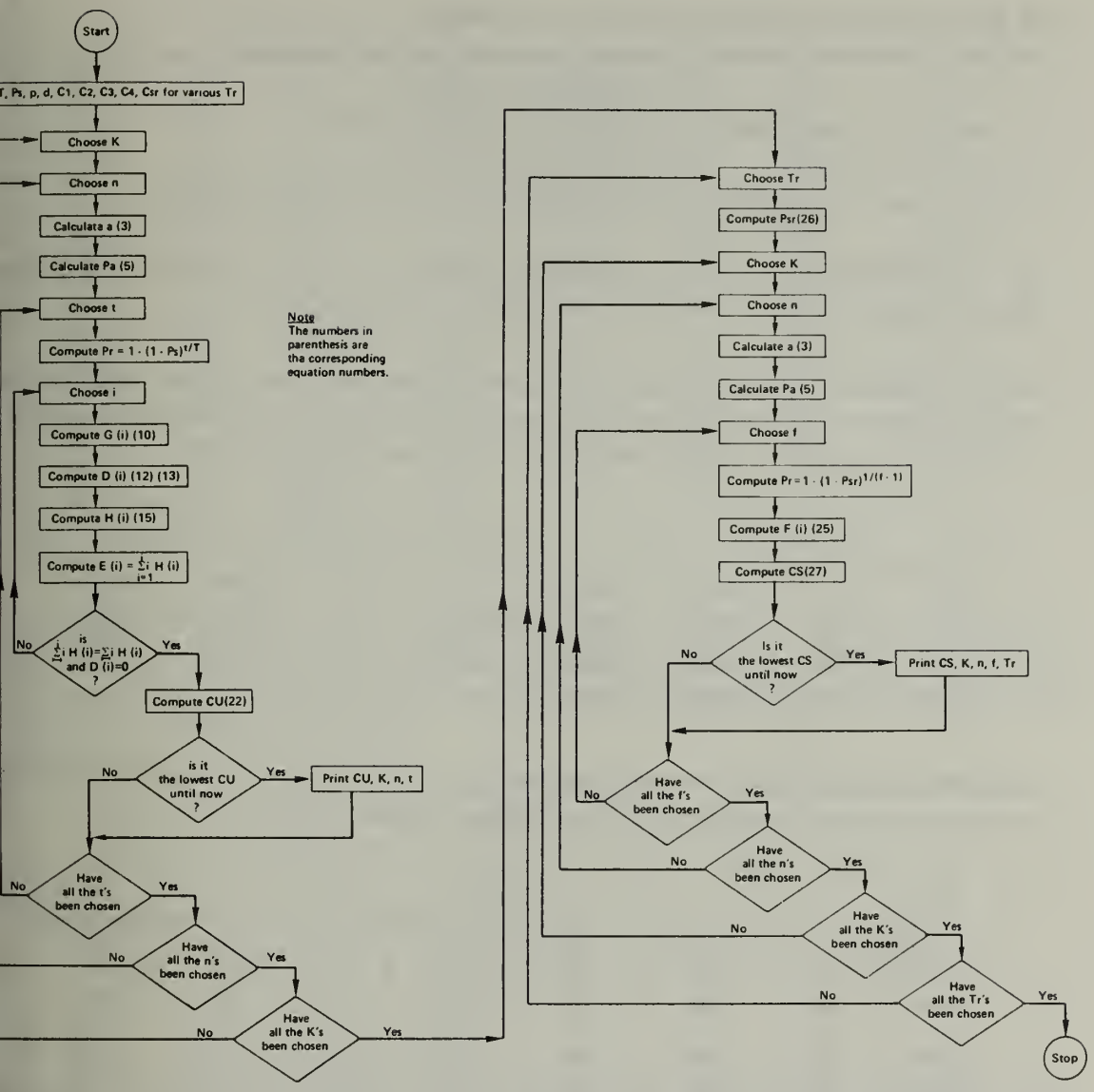


FIGURE 3. Flow chart for computing minimal cost per unit of time for scheduled and unscheduled setups.

When the control of the process by means of control charts has been considered, the following estimates were made:

(1) Cost of sampling one item: $C_1 = \$1$/item.

(2) Cost of one item (Labor and material) = \$5/unit.

(3) Cost of not detecting a change of 4 percent or more in the fraction defectives for 8 hours: 00 units produced per 8 hours @ \$5 unit $\times 0.04 = C_2 = \$2{,}000/8$ hours.

(4) Cost of each false indication = Cost of 5.2 minutes of loss of production + Cost for engaging rs in investigating the correctness of the setup:

$$\frac{\text{minutes} \times 10{,}000 \text{ units produced per 8 hours}}{480 \text{ minutes per 8 hours}} \times \$5/\text{unit} + \$50 = C_3 = \$600/\text{false indication}.$$

(5) Cost of setting up each detected shift: $C_4 = C_3 = \$600$.

(6) Cost of Scheduled Setup at the end or at the start of the 8 hour shift:

$$C_5 = \$20/\text{Setup for } T_r = T = 8.$$

(7) Cost of Scheduled Setup in midday or of any Unscheduled Setup:

$$C_{5r} = \$50/\text{Setup for } T_r \neq T = 8.$$

Computer search along the developed model has provided the following results:

Min $CS = \$7.26$/hour attained at $T_r = 16$ hours (scheduled setup every second shift), $f = 2$ sam per day (a sample each $[8/(2+1)]60 = 160$ minutes), $n = 3$, and $k = 3.5$.

Min $CU = \$125$/hour attained at infinity, i.e., if unscheduled setup policy is used, then no samp should be followed, and thus the control by means of control charts would not be economical. result has been found by searching for values of CU at very high values of t:

CU	K	n	t	CU	K	n	t
125.26	1.0	2,000.0	10,000.0	125.04	1.0	2,000.0	60,000.0
125.13	1.0	2,000.0	20,000.0	125.04	1.0	2,000.0	70,000.0
125.09	1.0	2,000.0	30,000.0	125.03	1.0	2,000.0	80,000.0
125.07	1.0	2,000.0	40,000.0	125.03	1.0	2,000.0	90,000.0
125.05	1.0	2,000.0	50,000.0				

It should be noted that if a nonoptimal unscheduled policy is used, the cost is considerably hig sometimes in order of magnitudes, as the following examples show:

CU	K	n	t	CU	K	n	t
1,241.24	3.50	9.00	0.20	315.85	3.50	23.00	15.80
1,247.30	3.50	9.00	0.60	353.06	3.50	25.00	0.20
1,248.08	3.50	9.00	1.00	287.78	3.50	25.00	0.60
1,248.37	3.50	9.00	1.40	275.72	3.50	25.00	1.00
1,248.52	3.50	9.00	1.80	270.68	3.50	25.00	1.40
1,248.60	3.50	9.00	2.20	267.92	3.50	25.00	1.80
1.248.66	3.50	9.00	2.60	266.17	3.50	25.00	2.20
1,248.70	3.50	9.00	3.00	264.96	3.50	25.00	2.60
1,248.73	3.50	9.00	3.40	264.08	3.50	25.00	3.00
1,248.75	3.50	9.00	3.80	263.40	3.50	25.00	3.40
1,248.76	3.50	9.00	4.20	262.87	3.50	25.00	3.80
1,248.77	3.50	9.00	4.60	262.44	3.50	25.00	4.20
1,248.78	3.50	9.00	5.00	262.08	3.50	25.00	4.60
1,248.79	3.50	9.00	5.40	261.79	3.50	25.00	5.00
1,248.79	3.50	9.00	5.80	261.53	3.50	25.00	5.40
1,248.79	3.50	9.00	6.20	261.31	3.50	25.00	5.80
1,248.79	3.50	9.00	6.60	261.12	3.50	25.00	6.20
1,248.80	3.50	9.00	7.00	260.95	3.50	25.00	6.60
1.248.80	3.50	9.00	7.40	260.79	3.50	25.00	7.00
1,248.80	3.50	9.00	7.80	260.66	3.50	25.00	7.40
1,248.80	3.50	9.00	8.20	260.54	3.50	25.00	7.80
1,248.79	3.50	9.00	8.60	260.43	3.50	25.00	8.20
1,248.79	3.50	9.00	9.00	260.33	3.50	25.00	8.60
1,248.79	3.50	9.00	9.40	260.24	3.50	25.00	9.00

CU	*K*	*n*	*t*	*CU*	*K*	*n*	*t*
1,248.79	3.50	9.00	9.80	260.15	3.50	25.00	9.40
1,248.79	3.50	9.00	10.20	260.07	3.50	25.00	9.80
1,248.78	3.50	9.00	10.60	260.00	3.50	25.00	10.20
1,248.78	3.50	9.00	11.00	259.93	3.50	25.00	10.60
1,248.78	3.50	9.00	11.40	259.87	3.50	25.00	11.00
1,248.77	3.50	9.00	11.80	259.81	3.50	25.00	11.40
1,248.77	3.50	9.00	12.20	259.76	3.50	25.00	11.80
1,248.77	3.50	9.00	12.60	259.71	3.50	25.00	12.20
1,248.76	3.50	9.00	13.00	259.66	3.50	25.00	12.60
1,248.76	3.50	9.00	13.40	259.61	3.50	25.00	13.00
1,248.76	3.50	9.00	13.80	259.57	3.50	25.00	13.40
1,248.75	3.50	9.00	14.20	259.53	3.50	25.00	13.80
1,248.75	3.50	9.00	14.60	259.49	3.50	25.00	14.20
1,248.75	3.50	9.00	15.00	259.46	3.50	25.00	14.60
1,248.74	3.50	9.00	15.40	259.42	3.50	25.00	15.00
1,248.74	3.50	9.00	15.80	259.39	3.50	25.00	15.40
315.92	3.50	23.00	15.00	259.36	3.50	25.00	15.80
315.89	3.50	23.00	15.40				

t is interesting to note that, for the unscheduled setup policy with the optimal parameters of the luled policy (i.e. for $t = 2\frac{2}{3}$ hours, $n = 3$, and $K = 3.5$), the cost reaches skyrocketing values. On ther hand, the cost of scheduled policy with nonoptimal decision variables also changes in a very range as the following table exhibits:

CS	*K*	*n*	*f*	T_r	*CS*	*K*	*n*	*f*	T_r
736.0	1.0	5	2	0.5	38.5	2	6	2	4
101.0	1.0	5	2	2.0	13.4	2	6	2	8
344.0	3.5	3	2	0.5	10.4	3	4	2	12
107.0	3.5	3	2	2.0	13.2	3	4	2	8

Analysis of the costs for unscheduled policy shows (for the given set of cost and process parame- that increase in K has a very minor effect on the total cost per unit of time for high values of T_r. decreases, the influence of K is more pronounced by decreasing CS with the increase of K. In all ns of T_r the decrease of CS by increasing K from 1.0 to 1.5 is significantly greater than by increas- from 1.5 to 3.5. The increase of T_r from 1 to 8 rapidly reduces CS for all values of K, while further ase in T_r slowly reduces CS to the minimal value. Above the optimal T_r of 16, there is only a slight ge in CS. Surprisingly, for all combinations of T_r and K in the present example, the minimal cost was obtained always for $f = 2$ and for a value of n which depended only on the value of K:

K	Optimal *n*	*K*	Optimal *n*	*K*	Optimal *n*
1.0	5	2.0	6	3.0	4
1.5	3	2.5	5	3.5	3

CONCLUSION

A general cost model has been developed for the selection of the optimal setup policy for a ma facturing process which uses a fraction-defective control chart to control the current production. selection is between an unscheduled setup policy and the set of all the possible scheduled setup poli with different fixed setup intervals. The selection criteria is the minimum expected cost per uni time, which considers the cost of sampling, the cost of not detecting an unwanted process behav the cost of false indications of unwanted process behavior, and the cost of scheduled and unschedu setups.

For any particular example, given the relevant cost data and the managerial value judgm data, the optimal unscheduled setup policy can be determined, thus providing the optimal deci variables of sample size, extent of control limits from process average, and fixed time interval betw subsequent samples, as well as the resulting expected total cost per unit of time. Simultaneously optimal scheduled setup policy can be determined, providing the corresponding optimal sample s extent of control limits from process average, the fixed time interval between scheduled setups, frequency of the equally spaced samples in this interval, and the resulting expected total cost per of time are given. The optimal unscheduled and the optimal scheduled setup policies can be deri using the appropriate expected total cost per unit of time formulae, Equations (22) and (28), which be screened by computer enumeration—according to the given flow chart—to determine the requi optimal values of the decision parameters. The setup policy having the lowest expected cost per of time in its optimal condition will be chosen as the optimum optimorum. It has to be stressed that, to the sensitivity of total cost per unit of time function to the cost and process parameters, i impossible to predict in the given situation the optimal values of the decision parameters and favorable setup policy from the presented numerical example, and each case should be evalua separately. Furthermore, the numerical example shows clearly that reliance on traditional rule thumb for selecting the spread of the control limits and the sample size in both setup situations m lead to significant cost increases above the attainable optimal solution.

ACKNOWLEDGMENT

The authors are indebted to Yoram Alperovitch, Tel-Aviv University, for his contribution to early parts of this paper and to an unknown referee for his helpful suggestions.

REFERENCES

[1] Baker, K. R., "Two Process Models in the Economic Design of an $\bar{X}$ Chart," AIIE Transactio *4* (1971).

[2] Duncan, A. J., "The Economic Design of $\bar{X}$ Charts Used to Maintain Current Control of a Proce J. of the Am. Stat. Assoc. *51*, 228–252 (1956).

[3] Duncan, A. J., "The Economic Design of $\bar{X}$ Charts When There is a Multiplicity of Assign Causes," J. of the Am. Stat. Assoc., *66* (Mar. 1971).

[4] Gibra, I. N., "Economically Optimal Determination of the Parameters of an $\bar{X}$ Control Cha Management Science *17*, *9* (1971).

[5] Goel, A. L., S. D. Jain, and S. M. Wu, "An Algorithm for the Determination of the Econ Design of $\bar{X}$-Charts Based on Duncan's Model," J. of the Am. Stat. Assoc. *63* (1968).

lall, R. I. and S. Eilon, "Controlling Production Processes Which are Subject to Linear Trends," Operational Research Quarterly *14*, *3*, 279–289 (Sept. 1963).

Inappenberger, H. A. and A. H. Grandage, "Minimum Cost Quality Control Tests," AIIE Transactions, *1*, *1* (1969).

adany, S. P., "Optimal Use of Control Charts for Controlling Current Production," Management Science *19*, *7*, 763–772 (Mar. 1973).

Iontgomery, D. C. and P. J. Klatt, "Economic Design of T^2 Control Charts to Maintain Current Control of a Process," Management Science, *19*, *1* (Sept. 1972).

OPTIMAL FLOWSHOP SCHEDULES WITH NO INTERMEDIATE STORAGE SPACE

Jatinder N. D. Gupta

Management Information Systems Department
U.S. Postal Service
Washington, DC

ABSTRACT

In this paper the problem of finding an optimal schedule for the n-job, M-machine flowshop scheduling problem is considered when there is no intermediate space to hold partially completed jobs and the objective function is to minimize the weighted sum of idle times on all machines. By assuming that jobs are processed as early as possible, the problem is modeled as a traveling salesman problem and solved by known solution techniques for the traveling salesman problem. A sample problem is solved and a special case, one involving only two machines, is discussed.

TRODUCTION

The problem of scheduling n jobs on M machines in a flowshop has been extensively considered iterature, and several solution procedures are available to solve the flowshop scheduling problems er several restrictive assumptions (see References [1, 3, 4, and 7] for recent reviews). One of these umptions, which is inherent in the classical formulation of the problem, is the availability of enough rmediate storage space to hold any number of partially processed jobs in case these jobs cannot processed further because of the preoccupation of other machines. However, in many practical ations, a job must be processed uninterrupted on all machines once it starts processing on the machine. In still other cases, no intermediate space is available to hold the partially processed jobs. ile considering these two cases, it is easy to visualize the inadequacy of the existing flowshop eduling techniques to solve this problem.

Recently, Reddi and Ramamoorthy [8] and Wismer [9] proposed a traveling salesman modeling he problem under consideration. However, neither of these two papers provides a theoretical ndation of their methods. This paper provides a theoretical basis of these developments and gests some extensions of the results of Reddi and Ramamoorthy [8] and Wismer [9]. Computational parison of the methods is provided which illustrates that the computational effort for problem nulation and solution by Reddi and Ramamoorthy's method is less than that by Wismer's method.

OBLEM DEFINITION AND ANALYSIS

The flowshop scheduling problem with no intermediate storage space may be defined as follows: "Given n jobs to be processed on M machines in the same technological order, the process time of a on machine m is t_{am}, ($a = 1, 2, \ldots, n$; $m = 1, 2, \ldots, M$), and a job once started is processed

through all the machines without interruptions or delay. Assuming that jobs are processed as soo possible, it is required to find the order (schedule) of processing these jobs on each of the M mach so as to minimize the weighted sum of the idle times on all machines."

The above definition of the problem gives rise to a possible $(n!)^M$ schedules. However, the cond of processing jobs without any intermediate storage space restricts the number of feasible sched to $n!$ as shown below:

THEOREM 1. For the flowshop scheduling problem with no intermediate storage spac feasible schedule has the same sequence of jobs on all machines.

PROOF. Consider two consecutive machines m_1 and m_2. Let job b immediately follow job machine m_1 and job a immediately follow job b on machine m_2. In such case, job a cannot start machine m_2 unless job b has completed its processing on machines m_1 and m_2. Therefore, afte completion on machine m_1, job a must wait for at least $t_{bm_1}+t_{bm_2}$ time units before starting on mac m_2. Since the processing times are non-negative, this waiting time can be zero if and only if j has zero processing times on all machines (m_1 and m_2 are arbitrary). Therefore, in the sche considered, job a cannot be processed uninterrupted on all machines. This argument can be conti further by allowing the processing of some other jobs between b and a on machine m_2. Since m_1, m_2 and b are all arbitrary, it follows that a feasible schedule to our problem has the same sequen jobs on all machines. Hence, the proof of Theorem 1. The results of Theorem 1 are quite signifi for without the restriction of no intermediate storage, the schedule obtained by assuming the passing" condition may not be optimal [1, 4, 7].

In a flowshop, the technological ordering of all jobs is the same and the numbering of mach is arbitrary. Hence, the machines can be numbered such that jobs are processed on machine first, machine two second, ..., and machine M last. With this reindexing of machines, it is to observe that if jobs do not wait at any intermediate machines, then they must be delayed at mac one and machine one must be kept idle between the processing of jobs. Consider a pair of jobs (where job a immediately precedes job b. Let $D(a, b)$ be the idle time on machine one between completion of job a and the start of job b. Further consider a partial schedule σab obtained by augn ing job pair (a, b) to partial schedule σ, i.e. σ precedes the job pair (a, b). Then, following the phy restrictions of the problem (nonsimultaneous processing of two or more jobs on the same machine nonsimultaneous processing of a single job on more than one machine), the completion time of pa schedule $\sigma\, ab$ on machine m, $T(\sigma ab, m)$, is given by the following recursive relation:

$$T(\sigma ab, m)=\max\,[T(\sigma ab,\ m-1);\ T(\sigma a,\ m)]+t_{bm};\ 2\leq m\leq M \tag{1}$$

$$T(\sigma ab, 1)=T(\sigma a, 1)+D(a, b)+t_{b1} \tag{2}$$

where for a null schedule ϕ

$$T(\phi,\ m)=D(0,\ b)=0\ \forall\ b \text{ and } m.$$

Let the idle time between two consecutive jobs a and b on machine m be $I_m(a, b)$ and the weigh tached to machine m be w_m. Since jobs are processed uninterrupted, $D(a, b)$ is such that rela (1) and (2) give:

$$T(\sigma ab, m) = T(\sigma ab, m-1) + t_{bm} = T(\sigma a, 1) + D(a, b) + \sum_{s=1}^{m} t_{bs}.$$

erefore:

$$I_m(a, b) = T(\sigma ab, m) - T(\sigma a, m) - t_{bm}, \text{ or}$$

$$I_m(a, b) = D(a, b) + \sum_{s=1}^{m-1} t_{bs} - \sum_{s=2}^{m} t_{as}.$$

$f(\sigma ab)$ is the value of the objective function for the partial schedule σab, then:

$$f(\sigma ab) = f(\sigma a) + \sum_{m=2}^{m} w_m I_m(a, b),$$

ce machine one has fixed idle time. However, for Equation (4), we must assume an initial boundary ndition, i.e. $f(\phi) = 0$.

We can now define the flowshop scheduling problem as one of minimizing $f(\sigma ab)$ where σ ranges er all possible permutations of $(n-2)$ jobs not including jobs a and b, and jobs a and b range over possible values 1 through n, that are not equal to each other.

The objective function in the above formulation represents several practical cases. Thus, for ample, if $w_m = 0$, $\forall m \leq M-1$ and $w_M = 1$, the problem reduces to the minimization of make-span, ined as the total throughput time required to process all jobs on all machines. If, however, the chines are rented and w_m is the rent of machine m per unit time, then the use of Equation (4) above l minimize the payment of rent for the total shop (as it will minimize the rent for idle machines).

EVELOPMENTS

The quantity $D(a, b)$ in Equations (2) and (3) above can be determined independent of schedule s shown below:

THEOREM 2. If jobs are processed as soon as possible and no intermediate space is available, n:

$$D(a, b) = \max_{2 \leq m \leq M} \left[\sum_{s=2}^{m} t_{as} - \sum_{s=1}^{m-1} t_{bs}, 0 \right].$$

PROOF. By definition, a job once started is processed without interruptions on all machines. nce, for any machine m:

$$T(\sigma a, m) = T(\sigma a, 1) + \sum_{s=2}^{m} t_{as} \text{ and}$$

$$T(\sigma ab, m-1) = T(\sigma ab, 1) + \sum_{s=2}^{m} t_{bs}.$$

In order to satisfy Equation (1) above:

$$T(\sigma ab, m-1) - T(\sigma a, m) \geq 0 \forall 2 \leq m \leq M. \tag{8}$$

Substituting the values of $T(\sigma a, m)$ and $T(\sigma ab, m-1)$ from Equations (6) and (7) in Relation (8) simplifying the resulting expression with the help of Equation (2) we get:

$$D(a, b) \geq \sum_{s=2}^{m} t_{as} - \sum_{s=1}^{m-1} t_{bs}, \forall 2 \leq m \leq M. \tag{9}$$

However, the noninterference of jobs at machine one requires that $D(a, b) \geq 0$. Combining this dition with the assumption of processing jobs as early as possible and Relation (9), the result Hence the validity of Theorem 2.

The results of Theorem 2 show that $D(a, b)$ is independent of the schedule σ. Therefore, I_m (in Equation (3) is independent of the partial schedule σ. Having obtained the expression for I_m (it now remains to find the schedule $(a_1, a_2, \ldots, a_n)$ such that the expression

$$\sum_{i=0}^{n-1} \left[\sum_{m=2}^{M} w_m I_m (a_i, a_{i+1}) \right]$$

is minimum where $a_0 = 0$ with $t_{a_0 m} = 0 \forall m \leq M$. It is convenient to model this problem as a tra salesman problem and to apply known solution techniques to solve the problem. Of course, the ar of computational time required will depend on the computational efficiency features of the tech used to solve the traveling salesman problem.

The traveling salesman problem consists of finding a tour $(a_1, a_2, \ldots, a_n, a_1)$ of a salesm minimize the total cost of travel. If the travel cost from city i to city j is C_{ij}, then the traveling sale problem is one of finding a sequence $(a_1, a_2, \ldots, a_n)$ such that the total cost, TC, given by

$$TC = \sum_{i=1}^{n-1} C_{a_i a_{i+1}} + C_{a_n a_1} \tag{10}$$

is minimum. Since it is not possible for a given city to follow itself in the sequence, $C_{ii} = \infty \forall$

In order to model the flowshop scheduling problem as a traveling salesman problem, we defi $(n+1)$ city problem where C_{ab} is as follows:

$$\left.\begin{aligned} C_{0a} &= \sum_{m=2}^{M} w_m I_m (0, a) = \sum_{m=2}^{M} w_m \left(\sum_{s=1}^{m-1} t_{as} \right) \\ C_{ab} &= \sum_{m=2}^{M} w_m I_m (a, b), \text{ and} \\ C_{a0} &= \sum_{m=2}^{M} w_m I_m (a, 0) = 0 \\ C_{aa} &= \infty \end{aligned}\right\} \quad Va, b \leq n. \tag{11}$$

state the following theorem:

THEOREM 3. The schedule $(a_1, a_2, \ldots, a_n)$ is optimal to the flowshop scheduling problem no intermediate storage space if and only if the tour $(0, a_1, a_2, \ldots, a_n, 0)$ is optimal for the 1) city problem with city 0 as the starting city and the cost matrix defined by Equation (11) above.

PROOF. By rewriting Equation (10) for total cost of the traveling salesman problem, it is seen

$$TC = C_{0a_1} + \sum_{i=1}^{n-1} C_{a_i a_{i+1}} + C_{a_n 0}.$$

plifying the above expression with the help of Equation (11), we get:

$$TC = \sum_{m=2}^{M} w_m \left(\sum_{s=1}^{m-1} t_{as} \right) + \sum_{i=1}^{n-1} w_m I\,(a_i, a_{i+1}).$$

total travel cost expression (12) for the tour $(0, a_1, a_2, \ldots, a_n, 0)$ is also the total cost equation he schedule $(a_1, a_2, \ldots, a_n)$ for the flowshop scheduling problem under consideration. efore, an optimal schedule for the flowshop scheduling problem is an optimal tour for the traveling man problem (and vice versa) provided the traveling salesman is stationed at city 0 and returns to) after visiting all the other n cities.

Several methods are available to solve the traveling salesman problem [2, 6]. Depending on the lem size, a suitable method can be selected and used to seek the solution to the problem.

UMERICAL ILLUSTRATION

As an illustration, consider the four-job, five-machine problem with process times in Table 1.

TABLE 1. *Process Times of Each Job on Each Machine*

a \ m	1	2	3	4	5
1	4	3	7	2	8
2	3	7	2	8	5
3	1	2	4	3	7
4	3	4	3	7	2

assume that $w_m = 0 \forall m \leq M-1$ and $w_M = 1$.

The $D(a, b)$ values are given by Equation (5) above. In order to illustrate, consider the deter- tion of $D(1, 2)$. From Equation (5),

$$D(1, 2) = \max_{2 \leq m \leq 5} \left[\sum_{s=2}^{m} t_{1s} - \sum_{s=1}^{m-1} t_{2s}, 0 \right]$$

$$= \max [3-3, 3+7-3-7, 3+7+2-3-7-2, 3+7+2+8-3-7-2-8, 0]$$

$$= 0.$$

Continued use of Equation [5] yields the delay matrix in Table 2.

TABLE 2. *Delay Matrix*

a \ b	1	2	3	4
1	—	0	10	3
2	6	—	12	7
3	0	0	—	0
4	0	2	7	—

Using Equations (3) and (11), the travel cost matrix in Table 3 is obtained.

TABLE 3. *Travel Cost Matrix*

a \ b	(0)	1	2	3	4
(0)	∞	16	20	10	17
1	0	∞	0	0	0
2	0	0	∞	0	2
3	0	0	4	∞	1
4	0	0	6	1	∞

Solving the above traveling salesman problem by any known technique [2, 6], the optimal is (0, 3, 4, 1, 2, 0) with a total cost of 11 units. Therefore, the optimal schedule for our flowshop pro is (3, 4, 1, 2) and has 11 units of idle time on the last (fifth) machine.

MPUTATIONAL RESULTS

Considerable experimentation was conducted to investigate the efficiency of the proposed problem ıulation (which is a generalized form of the one described by Reddi and Ramamoorthy) and Wis-'s formulation. For this purpose, the two methods were programmed for IBM 360/65 computer in RTRAN. The comparison of the methods (or alternate formulations) was limited to the formulation he flowshop scheduling problem as the traveling salesman problem, since computational effort ond this point is a function of the techniques used to solve the corresponding traveling salesman blem. In order to carry out the experimental investigation, 300 problems with the number of jobs ging from 10 through 25 and the number of machines ranging from six through 25 were generated formulated as traveling salesman problems by the proposed method and Wismer's method. The cess times of the jobs in the above problems were randomly generated from a discrete rectangular ribution. They ranged from 000 through 999 and all were integers. Further, the comparison was ricted to the formulation of problems with objective functions as the minimization of make-span. le 4 gives the computation times for various problems by both methods. (These computation times ot include the times for reading the data or printing the corresponding traveling salesman problem.)

TABLE 4. *Comparative Evaluation of Proposed and Wismer's Methods*

$n \times M$	Number of problems	Computations time (sec) by proposed method		Computation time (sec) by Wismer's method	
		Average	Range	Average	Range
10 × 6	20	0.0165	0.0160–0.0170	0.0300	0.0160–0.0340
10 × 10	20	0.0222	0.0160–0.0340	0.0525	0.0160–0.0830
10 × 25	20	0.0475	0.0330–0.0500	0.0859	0.0500–0.1170
15 × 6	20	0.0310	0.0170–0.0340	0.0724	0.0500–0.0840
15 × 8	20	0.0408	0.0160–0.0670	0.0909	0.0670–0.1170
15 × 10	20	0.0508	0.0330–0.0670	0.1049	0.0830–0.1330
15 × 25	20	0.1017	0.0500–0.1340	0.1966	0.1660–0.2330
20 × 6	20	0.0567	0.0330–0.0670	0.1267	0.1000–0.1340
20 × 8	20	0.0742	0.0500–0.1000	0.1576	0.1330–0.1840
20 × 10	20	0.0867	0.0670–0.1000	0.1809	0.1000–0.2170
20 × 25	20	0.1950	0.1660–0.2340	0.3590	0.3330–0.4000
25 × 6	20	0.0850	0.0670–0.1170	0.1850	0.1500–0.2000
25 × 8	20	0.1040	0.0830–0.1170	0.2300	0.2000–0.2500
25 × 10	20	0.1267	0.1000–0.1500	0.2700	0.2170–0.3000
25 × 25	20	0.2916	0.2660–0.3170	0.5300	0.5000–0.5670

From the results of Table 4, it is obvious that the proposed method (formulation) is comparatively e efficient than Wismer's method, as the average time for problem formulation as well as the range he proposed method is less than that by Wismer's method.

TWO-MACHINE PROBLEMS

If the number of machines is $M = 2$, then a method of solving a one-state variable machir quencing problem developed by Gilmore and Gomory [5] can be used. Gilmore and Gomory consi the problem of sequencing n jobs on a single machine whose state is described by a real varial If job i requires a starting state A_i and a finish state B_i, the changeover cost from i to j is define

$$C_{ij} = \begin{cases} \int_{B_i}^{A_j} f(x)\,dx & \text{if } A_j \geq B_i \\ \int_{A_j}^{B_i} g(x)\,dx & \text{if } A_j < B_i \end{cases}$$

where functions $f(x)$ and $g(x)$ are integrable and $g(x) + f(x) \geq 0$. By setting $A_i = t_{i1}, B_i = t_{i2}, f(x) =$ and $g(x) = 0$, we see that the above problem represents the two-machine flowshop scheduling pro with no intermediate storage space. Therefore, the method proposed by Gilmore and Gomory solve the two-machine case of our problem. As shown by these authors, the method is much efficient than any other technique for solving the traveling salesman problem. (For details of the me see reference [5].)

CONCLUSIONS AND RECOMMENDATIONS

This paper has described a method for solving the flowshop scheduling problem when no mediate space is available to hold partially processed jobs. However, there are several allied prob that still need solution procedures. For example, if the maximum number of jobs that can be ma wait before machine m is a finite positive number F_m, then the proposed method will not solve the lem. Similarly, if the objective function is either the minimization of mean flow time or the mini tion of the penalty of late jobs, the proposed solution technique fails to generate an optimal sche Developments of solution procedures to solve these problems will greatly facilitate the practical a cations of quantitative methods for managerial decisions.

REFERENCES

[1] Bakshi, M. K. and S. R. Arora, "The Sequencing Problem," Management Science, *16*, B247– (1969).

[2] Bellmore, M. and G. L. Nemhauser, "The Traveling Salesman Problem, A Survey," Opera Research, *16*, 538–558 (1968).

[3] Day, J. E. and M. P. Hottenstien, "Review of Sequencing Research," Nav. Res. Log. Quar 11–39 (1970).

[4] Elmaghraby, S. E., "The Machine Sequencing Problem—Review and Extensions," Nav. Res. Quart. *15*, 205–232 (1968).

[5] Gilmore, P. C. and R. E. Gomory, "Sequencing a One State-Variable Machine: A Solvable of the Traveling Salesman Problem," Operations Research *12*, 655–679 (1964).

[6] Gupta, J. N. D., "Traveling Salesman Problem: A Survey of Theoretical Developments and A cations," Opsearch (India), *5*, 181–192 (1968).

Gupta, J. N. D., "*M*-Stage Scheduling Problem: A Critical Appraisal," The International Journal of Production Research, *9*, 267–281 (1971).

Reddi, S. S. and C. V. Ramamoorthy, "On the Flowshop Sequencing Problem With No Wait in Process," Operational Research Quarterly (1972).

Wismer, D. A., "Solution of the Flowshop Scheduling Problem With No Intermediate Queues," Operations Research *20*, 689–697 (1972).

LONGITUDINAL MANPOWER PLANNING MODELS*

Richard C. Grinold, Kneale T. Marshall and Robert M. Oliver

Operations Research Center
University of California, Berkeley

ABSTRACT

A manpower planning model is presented that exploits the longitudinal stability of manpower cohorts. The manpower ning process is described. An infinite horizon linear program for calculating minimum cost manpower input plans is pre- d and found to have a straightforward solution in a great many cases and to yield an easily implemented approximation nique in other cases.

INTRODUCTION

This paper formulates manpower planning models for a system consisting of many skill categories; particular application is to the enlisted force in the U.S. Navy. Interactive computer models of e theoretical formulations have been developed and implemented to aid decisionmakers who ı to test the effects of alternative manpower policies on staffing requirements and future man- er budgets. These interactive manpower planning models can be used to:

(1) predict the manpower requirements that will be fulfilled by the current stock of manpower.
(2) calculate unfulfilled requirements and the new inputs necessary to meet them.
(3) identify bottlenecks in the manpower planning process.
(4) assist in preparation of future manpower budgets.
(5) simulate the effects of manpower policy changes on future manpower needs.
(6) relate alternate personnel retention and performance assumptions to the need for future inputs.
(7) calculate minimum cost input schedules when lower bounds on requirements are given.

ails of the interactive models and data utilized are treated in [4]. The models presented in this er are of general interest, but we shall use the motivating context of the U.S. Navy enlisted force escribe the model.

Individuals in any skill category can be identified by several characteristics. Examples are rank, ry, number of years of experience in the skill category, length of service in the Navy, and personal butes such as age and measures of performance. The models presented in this paper are designed ssist in preparing manpower budgets and meeting aggregate strength requirements. For these poses, we have chosen to identify individuals in the enlisted force according to skill category and th of service (LOS) in the Navy.

---

*This research has been supported by the Office of Naval Research under Contract N00014-69-A-0200-1055 with the ersity of California. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

Section 2 of the paper describes the underlying manpower flow models. The models are ba on the assumption of longitudinal stability in the service lifetimes of different manpower cohorts. show that the accession schedule that exactly meets requirements is bound by solving a set of lo triangular system of linear equations. In Section 3, we present an infinite horizon linear program the calculation of future manpower inputs and strength levels with the objective of minimizing counted costs. In Section 4, we derive readily verifiable conditions on the inputs to the infinite hori problem that guarantee that certain easily calculated policies will be optimal. Finally, in Sectio we examine cases where simple policies are not optimal and describe a finite linear program that approximate solutions of the infinite horizon program.

The models presented in this paper examine the relationships between three factors: (i) the rent manpower situation as described by the LOS distributions of different skill ratings, (ii) the vivor fractions that will determine the longitudinal behavior of manpower cohorts and (iii) the manpo requirements for future times. The size of our models and the type of calculations performed al policymakers to quickly analyze the impact of various assumptions and policies.

2. THE UNDERLYING COHORT FLOW MODELS

We consider an organization which is divided into many skill categories where movement betw categories typically involves retraining of the individual. In this paper we consider each skill cate separately; a subsequent paper will discuss transfers and other interactions between skill catego

We idealize the evolution of the skill category by analyzing its changes at discrete points in t ($i = \ldots -2, -1, 0, 1, 2, \ldots$). We say that period i is the interval between times $(i-1)$ an it is a future period if $i > 1$, a past period if $i < 0$, and the current period if $i = 1$. In period i, a num of people x_i enter the skill category; that group is called *cohort* i, and x_i is the number of *accessi* in cohort i. Let α_{ij} be the fraction of the cohort entering in period i which is still present and availa to meet requirements at time $i+j$ ($j \geq 0$). Let z_k be the requirement in the skill category at tim and let $(m+1)$ be the maximum number of periods an individual is allowed to stay in the syst Thus, $\alpha_{ij} = 0$ if $j > m$. For some future time k, we have

$$z_k = x_k\alpha_{k,0} + x_{k-1}\alpha_{k-1,1} + \ldots + x_{k-m}\alpha_{k-m,m}. \tag{1}$$

Equation (1) simply says that the requirement at time k is made up of fractions of cohorts which sur from earlier periods. Thus, it is natural to call the α_{ij}'s the *survivor fractions* for the cohort which en at time i.

At time 0, the history of past accessions is given by the vector $(x_{-m}, x_{1-m}, \ldots, x_{-2}, x_{-1},$ The current inventory of people is given by $x_0\alpha_{0,,0} + x_{-1}\alpha_{-1,1} + \ldots + x_{-m}\alpha_{-m,m}$, and conta the remaining fractions of past inputs. This quantity is called the *current legacy*, say y_0. In fu period k, the legacy y_k from past inputs up to and including period 0 will be

$$\begin{aligned} y_k &= x_0\alpha_{0,k} + x_{-1}\alpha_{-1,k+1} + \ldots + x_{k-m},\alpha_{-m,k+m} && \text{if } k \leq m, \\ &= 0 && \text{if } k > m. \end{aligned} \tag{2}$$

Suppose we have a planning horizon of T periods with requirements $z_1, z_2, \ldots, z_T$. From E tions (1) and (2), we see that future cohort sizes must satisfy

$$
\begin{aligned}
&\alpha_{1,0}x_1 && = z_1 - y_1,\\
&\alpha_{1,1}x_1 + \alpha_{2,0}x_2 && = z_2 - y_2,\\
&\cdots\\
&\cdots\\
&\alpha_{1,T-1}x_1 + \alpha_{2,T-2}x_2 + \cdots + \alpha_{T,0}x_T = z_T - y_T.\\
&\quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad \vdots\\
&\quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad \vdots
\end{aligned}
$$

e we have assumed

A1: requirements are met exactly.

Under A1, it is quite possible that for a given set of z_k's, y_k's and α_{ij}'s, some x_k could be negative. h a result would say that in order to exactly meet requirements in all periods $1, 2, \ldots, T$, it will necessary to remove people from the skill category in period k. In practice, this might be accomhed through retraining.

This section concentrates on the equality solution for several reasons. First, it is misleading to e the problem as if the accessions (the x_k) are the only variables which the decisionmaker can uence. The legacies (y_k), the requirements (z_k) and to some extent the survivor fractions (α_{ij}) all be changed or explicitly influenced by manpower policies. Second, plans that are eventually ommended will probably conform to the equality constraints since budget restrictions do not genlly allow for slack in the system. Third, we intend to use the models in this section to test the effects lternate policies on several objectives: (1) the departure of realistic requirements from ideal reements, (2) the smoothing of stocks and accessions, (3) the impact of policy changes on personnel finally (4) the retraining costs associated with switching personnel from one specialty to another. Section 3, we shall drop A1, treat the net requirements as lower bounds and look for cost minimizing ession schedules.

In the remainder of this paper, we make an important second assumption:

A2: the survivor fractions $\alpha_{i,j}$ are stationary from period to period. That is, $\alpha_{i,j} = \alpha_j$, independent of i and independent of x_i.

Under Assumption A2, Equation (3) simplifies to

$$
\begin{aligned}
&\alpha_0 x_1 && = z_1 - y_1,\\
&\alpha_1 x_1 + \alpha_0 x_2 && = z_2 - y_2,\\
&\cdots\\
&\alpha_{T-1}x_1 + \alpha_{T-2}x_2 + \cdots + \alpha_0 x_T = z_T - y_T.
\end{aligned}
$$

hese equations, the legacies are given by

$$
\begin{aligned}
y_1 &= \alpha_1 x_0 + \alpha_2 x_{-1} + \cdots + \alpha_m x_{1-m},\\
y_2 &= \alpha_2 x_0 + \alpha_3 x_{-1} + \cdots + \alpha_m x_{2-m},\\
&\cdots\\
y_T &= \alpha_T x_0 + \cdots + \alpha_m x_{T-m}.
\end{aligned}
$$

Equation (4) can be used in a number of ways. We have mentioned already that given the requ ments, legacies and survivor fractions (4) can be used to calculate new cohort input requirem for each period of the planning horizon T. Alternatively, given planned inputs over the next T peri the z_i's can be considered as the result of these inputs. Also, given requirements and planned inp the legacies which satisfy Equation (4) can be determined. These possibilities are discussed furt in [3].

To this point, we have motivated Equations (4) and (5) by considering z_i and y_i as stocks of pe and x_i as a flow of people per year. It is not necessary to define α, x y and z in that manner. We c also speak of stocks and flows of money or effective personnel or could use another unit of time. example, if c_j is the cost associated with an individual in the jth year of service, the total cost in i will be

$$c_0\alpha_0 x_i + c_1\alpha_1 x_{i-1} + \cdots + c_m\alpha_m x_{i-m}. \tag{6}$$

The cost legacy that will be incurred in future period i due to current period 0 manpower legacie

$$c_i\alpha_i x_0 + c_{i+1}\alpha_{i+1} x_1 + \cdots + c_m\alpha_m x_{i-m}. \tag{7}$$

3. MINIMIZING ACCESSION COSTS

The accession schedule that meets future manpower requirements z_t exactly is found by sol the equations

$$\sum_{j=1}^{t} \alpha_{t-j} x_j = z_t - y_t \qquad \text{for } t \geqq 1. \tag{8}$$

In this section, we shall relax the assumption that future manpower requirements are satis exactly. Instead we shall treat the variables z_t as lower bounds on the manpower level at time t. M over, we shall restrict the accessions x_t to be nonnegative. Thus, the equalities in (1) will be repla by inequalities ($\geqq$). This leaves us with an infinite system of linear inequalities that will, in gen have a large number of possible solutions. To obtain a single accession schedule in this case, we specify a performance criterion and then select the accession schedule that optimizes that crite

In the analysis that follows, we assume that the performance criterion is to minimize the pre worth of all future accession costs. This objective is obtained by discounting future costs to tod dollars and then summing over all future periods. We obtain the extremely simple and useful re that the optimal accession schedule is independent of the costs. In addition, we shall show in Sec 4 that there are many conditions of practical interest when the equality solution discussed in ea sections is indeed an optimal solution to the infinite-horizon program. Finally, in Section 5, we discuss several practical methods by which the infinite program can be truncated and approxim by a finite program with T rather than an infinite number of periods. The method of truncation should be used depends on the conditions the user wishes to assume for planning periods in the dis future.

There are several approaches to the solution of (8) when the restriction of exactly meeting the r power requirements is relaxed. In describing a cost minimization model in which the manpowe quirements are considered to be lower bounds, we assume that manpower requirements can in be modelled by such inequality restrictions.

The problem is to choose the nonnegative vector $(x_1, x_2, \ldots)$ that satisfies

$$\sum_{j=1}^{t} \alpha_{t-j} x_j \geqq z_t - y_t \qquad \text{for } t \geqq 1.$$
$$x_t \geqq 0$$

that it is possible to consider x_t as the number of accessions above some minimal level. Then ould be interpreted as the requirements left unsatisfied by the minimal accession schedule. A ble solution of (9) always exists if $\alpha_0 > 0$. It is given by

$$x_t = \text{Max}\left[0, \left(z_t - y_t - \sum_{j=1}^{t-1} \alpha_{t-j} x_j\right) \Big/ \alpha_0\right].$$

Let c_j for $j = 0, 1, 2, \ldots, m$ be the cost of training and support of an individual in the jth year of ce. The discounted cost of an accession is thus

$$c = \sum_{j=0}^{m} (\alpha_j c_j)\delta^j = \alpha_0 c_0 + \alpha_1 c_1 \delta + \alpha_2 c_2 \delta^2 + \ldots + \alpha_m c_m \delta^m$$

e δ is a discount factor less than one. The total discounted cost of an accession schedule $(x_1,$ . .) is therefore

$$\sum_{t=1}^{\infty} c x_t \delta^{t-1} = c x_1 + \delta c x_2 + \delta^2 c x_3 + \ldots ..$$

an accession in any period t costs $\delta^{t-1}c$, it is evident that the value of c will not affect the optimal ion. In fact, we can, without any loss in generality, use

$$\sum_{t=1}^{\infty} x_j \delta^{t-1} = x_1 + \delta x_2 + \delta^2 x_3 + \ldots$$

e objective to be minimized subject to the constraints in (9). The x_t that maximize (12) subject to onstraints (9) will also maximize (11) subject to (9). Thus, it is not necessary to have detailed in- ation about the cost structure to solve the optimal accession problem.

The dual of the infinite linear program (9), (10) is to find nonnegative variables $u_1, u_2, \ldots$ which

$$\text{Maximize } u_1(z_1 - y_1) + u_2(z_2 - y_2) + \ldots$$

ct to the inequality constraints

$$\sum_{j=0}^{m} u_{t+j} \alpha_j \leqq c\delta_{t-1}$$
$$u_t \geqq 0 \qquad \text{for all } t \geqq 1.$$

A feasible solution to (14) always exists and is given by

$$u_t = c\mu\delta^{t-1} \tag{15}$$

where

$$\mu = 1 \Big/ \left(\sum_{j=0}^{m} \alpha_j \delta^j\right).$$

This solution is strictly positive and satisfies each constraint of (14) as an equality.

The main results of linear programming do not carry over to infinite linear programs; the str duality theorem can fail [1], and in some cases even the weak duality theorem and the related st ciency condition (complementary slackness) can fail [3]. Fortunately, our problem has a great dea special structure and, under the reasonable Assumption A3, below, we shall be able to establis duality theorem and useful optimality conditions. These results will be used in Section 4 to char terize easily calculated optimal policies. The assumptions are

A3: (i) $$c = \sum_{j=0}^{m} (\alpha_j c_j)\delta^j > 0,$$

(ii) $$0 \leqslant \delta < 1.$$

(iii) $$\alpha_0 > 0,$$

(iv) $$\alpha_j \geqq 0, \qquad 1 \leqslant j \leqslant m,$$

(v) $$\sum_{t=1}^{\infty} z_t \delta^{t-1} < \infty.$$

THEOREM 1. Under Assumption A3, the primal infinite horizon program (9), (11) has an opti solution $\tilde{x}_t$, and the dual infinite horizon program (13), (14) has an optimal solution $\tilde{u}_t$; moreover,

$$\sum_{t=1}^{\infty} c\tilde{x}_t \delta^{t-1} = \sum_{t=1}^{\infty} \tilde{u}_t (z_t - y_t).$$

Finally, feasible solutions x_t of (9) and u_t of (14) are optimal if and only if they satisfy the complement slackness conditions

(16)

(i) $$u_t \left(\sum_{j=1}^{t} \alpha_{t-j} x_j - (z_t - y_t)\right) = 0$$

(ii) $$\left(\sum_{j=0}^{m} u_{t+j}\alpha_j - c\delta^{t-1}\right) x_t = 0$$

for all $t \geqq 1$.

The proof of Theorem 1 is presented in the Appendix.

The following section uses this result to characterize optimal policies, and Section 5 uses the cept of duality to find approximate solutions to the infinite primal problem.

OPTIMAL POLICIES

This section indicates conditions on the data that imply a simple form of the optimal solution to nfinite horizon linear program. In particular, we demonstrate that the solution

$$x_t = \text{Max}\left[0, \left(z_t - y_t - \sum_{j=1}^{t-1} \alpha_{t-j} x_j\right) \Big/ \alpha_0\right] \qquad t = 1, 2, \ldots$$

timal in several interesting cases. Notice that (17) defines the equality solution of equations

$$\sum_{j=1}^{t} \alpha_{t-j} x_j = z_t - y_t \qquad t = 1, 2, \ldots$$

at solution is nonnegative.

We shall first demonstrate conditions which imply that the unique equality solution of (18) is mal. Recall from (15) that $u_t = c\mu\delta^{t-1}$ is a feasible solution of the dual infinite horizon problem. e multiply the tth equation of (18) by those values of u_t, we obtain, after rearrangement,

$$\sum_{t=1}^{\infty} cx_t\delta^{t-1} = \sum_{t=1}^{T} u_t(z_t - y_t) = c\mu \sum_{t=1}^{\infty} (z_t - y_t)\delta^{t-1}.$$

term on the right is the value of a dual feasible solution, and is, therefore, a lower bound on the e of all primal feasible solutions. However, if x_t is nonnegative, it is primal feasible and attains ower bound; therefore, it is primal optimal.

We now present conditions on the problem data that guarantee that the equality solution is non- tive and, therefore, optimal. First note we must have $z_t - y_t \geqq 0$, and except for trivial cases, $y_t > 0$. Thus, we assume $z_t - y_t > 0$ for $t \geqslant 1$. To describe the other conditions, we introduce the on of a manpower continuation rate and a requirements growth rate. To define the continuation , let $\beta_0 = 1$ and for $1 \leqq j \leqq m$

$$\beta_j = \alpha_j / \alpha_{j-1}.$$

$j > m$, let $\beta_j = 0$. The continuation rate β_j is the fraction of those with length of service $j-1$ who inue in the system for at least one more period. Note that the survivor fractions can be defined rms of the continuation rates

$$\alpha_j = \alpha_0 \prod_{k=0}^{j} \beta_k.$$

growth rate ϕ_{t+1} is the increase in net requirements; thus, $\phi_{t+1}(z_t - y_t) = z_{t+1} - y_{t+1}$.

The net inflow in period $t+1$ is $\alpha_0 x_{t+1}$. In the equality solution of (18), this inflow has two func-

tions: first, to replace losses and second, to provide the growth increment. The loss in stock of acce sions that entered in periods 1 through t is given by

$$\sum_{j=1}^{t} (1-\beta_{t+1-j})\alpha_{t-j}x_j,$$

since $(1-\beta_{t+1-j})$ is the fraction of those with LOS between 1 $(j=t)$ and t $(j=1)$ that do *not* continue Therefore, $(1-\beta_{t+1-j})\alpha_{t-j}x_j$ is the number of individuals with LOS equal to $t-j$ that leave the system

The growth increment is given by

$$(\phi_{t+1}-1)(z_t-y_t) = (\phi_{t+1}-1)\sum_{j=1}^{t}\alpha_{t-j}x_j$$

where we have used (18) to eliminate z_t-y_t. The basic accounting relation for the equality solutio is, therefore,

$$\begin{array}{lll} \text{INFLOW} = & \text{GROWTH} & + \text{ REPLACEMENT} \\ \alpha_0 x_{t+1} & = (\phi_{t+1}-1)\sum_{j=1}^{t}\alpha_{t-j}x_j + & \sum_{j=1}^{t}(1-\beta_{t+1-j})\alpha_{t-j}x_j. \end{array}$$

When the terms on the right are combined, we obtain

$$\alpha_0 x_{t+1} = \sum_{j=1}^{t}(\phi_{t+1}-\beta_{t+1-j})\alpha_{t-j}x_j. \tag{20}$$

From (20), we can deduce that x_{t+1} will be nonnegative if

$$\phi_{t+1} \geqq \text{Max}\,[\beta_1, \beta_2, \ldots, \beta_t]. \tag{21}$$

Thus, (21) for all $t \geqq 0$ is sufficient for the equality solution of (18) to be nonnegative and optimal.

A necessary condition is simple to derive. Note that $x_1 = (z_1-y_1)/\alpha_0$, and for all $t \geqslant 2$

$$(z_t-y_t) - \frac{\alpha_t(z_1-y_1)}{\alpha_0} = \sum_{j=2}^{t}\alpha_{t-j}x_j. \tag{22}$$

If the equality solution x_t is nonnegative, then the right side of (22) is nonnegative, which in turn impli

$$\frac{(z_t-y_t)}{(z_1-y_1)} \geqq \frac{\alpha_t}{\alpha_0} \qquad \text{for all } t \geqq 2.$$

This can be rewritten as

$$\prod_{k=1}^{t}\phi_k \geqq \prod_{k=1}^{t}\beta_k \qquad \text{for all } t. \tag{23}$$

While $\phi_j \geq \beta_j$ implies (23), the converse is not true as it is quite possible that $\phi_j < \beta_j$ while $\phi_1\phi_2 \ldots \beta_1\beta_2 \ldots \beta_j$. Thus, the simple and local test of whether the growth rate in net requirements eds the continuation rate of individuals in the period 1 cohort lies somewhere between the neces- conditions of (23) and the more global sufficiency conditions in (21).

In several special cases, there are tighter and more easily verified conditions under which the lity solution is optimal.

First, if the β_j are nondecreasing, then $\phi_j \geq \beta_j$ implies $\phi_j \geq \beta_i$ for $i \leq j$. Thus the conditions β_j are necessary and sufficient for the equality solution to be nonnegative.

In a second case, if the α_j are nonincreasing, then $\beta_j \leq 1$ for all j. Moreover, nonincreasing α_j y that the legacy y_t is nonincreasing. If z_t is nondecreasing, it follows that $z_t - y_t$ is nondecreasing thus that $\phi_j \geq 1$ for all j. Therefore, we have optimality of the equality solution under the readily ied conditions z_t nondecreasing and α_j nonincreasing.

When the equality solution is nonnegative, the optimal dual variables are given by $c\mu\delta^{t-1}$. The cost of an optimal solution is

$$\left(\sum_{j=0}^{m} (\alpha_j c_j)\delta^j \Big/ \sum_{j=0}^{m} \alpha_j \delta^j\right) \sum_{t=1}^{\infty} (z_t - y_t)\delta^{t-1}.$$

formula has a reasonable interpretation in the case where $\alpha_0 = 1$ and the α_j are nonincreasing. $\alpha_j - \alpha_{j+1}$ be the probability that an individual's lifetime, i.e., maximum LOS, is equal to j. With stochastic interpretation of the survivor fractions, we define two random variables: T the individual's me, and K the total support cost of the individual. When δ is equal to 1, the term in brackets in is simply $E[K]/E[T]$. Thus, we can save by keeping $E[T]$ fixed and reducing costs. Notice, ever, that if we attempt to increase expected lifetime by changing the α_j, then the cost will change We can get a more accurate estimate of the impact of possible changes by rewriting the first term) with $\delta = 1$ and α_j expressed in terms of continuation rates

$$\left(\sum_{j=0}^{m} c_j \prod_{k=0}^{j} \beta_k\right) \Big/ \sum_{j=0}^{m} \prod_{k=0}^{j} \beta_k.$$

derivative of (25) with respect to β_l is

$$\frac{\sum_{j=l}^{m} \{c_j - E[K]/E[T]\}\alpha_j}{\beta_l E[T]}.$$

expression reinforces some intuitive feelings about the system. If the cost in the periods following greater than average, then an increase in β_l will surely increase costs. On the other hand, if all of downstream costs are less than average, then it will reduce costs when β_l is increased.

Notice that (17) corresponds to the equality solution of (18) when the equality solution is non- tive, i.e., optimal. Thus, (17) is an optimal solution in those cases. We indicate below that (17) otimal in more general cases. To demonstrate this, we use the complementary slackness con- ns (15)

$$\text{(i)} \qquad \tilde{u}_t \left(\sum_{j=1}^{t} \alpha_{t-j} \tilde{x}_j - (z_t - y_t) \right) = 0,$$

(26)

$$\text{(ii)} \qquad \left(\sum_{j=0}^{m} \tilde{u}_{t+j} \alpha_j - c\delta^{t-1} \right) \tilde{x}_t = 0.$$

Any feasible solution of the infinite problem must satisfy (17) as a greater than or equal to equality. Suppose $\tilde{x}_t$ is optimal but does not satisfy (17). There must be a smallest t such that

$$\tilde{x}_t > \text{Max} \left[0, \left(z_t - y_t - \sum_{j=1}^{t-1} \alpha_{t-j} \tilde{x}_j \right) \Big/ \alpha_0 \right].$$

It follows from (26), (i) that $\tilde{u}_t = 0$ and

$$\text{(27)} \qquad \sum_{j=0}^{m} \tilde{u}_{t+j} \alpha_j = \sum_{j=1}^{m} \tilde{u}_{t+j} \alpha_j = c\delta^{t-1}.$$

If we multiply (27) by δ and subtract from the $t+1$st dual constraint, we obtain

$$\text{(28)} \qquad \sum_{j=1}^{m} \tilde{u}_{t+j} (\alpha_{j-1} - \delta\alpha_j) + u_{t+1+m} \alpha_m \leqq 0.$$

If $\alpha_{j-1} > \delta\alpha_j$, then (28) can only be satisfied if $u_{t+j} = 0$ for $1 \leqq j \leqq m+1$. However, (27) indicates th cannot be true. Thus, our assumption that (17) fails to hold for an optimal solution leads to a contrad tion if $\alpha_{j-1} > \delta\alpha_j$.

The results of this section are summarized in Theorem 2.

THEOREM 2. If either of the conditions A4 or A5 holds, then (17) gives the optimal solution of t infinite horizon linear program.

A4: for all t

$$\phi_{t+1} \geqq \begin{cases} \text{Max } [\beta_1, \beta_2, \ldots, \beta_t] & \text{if } t < m \\ \text{Max } [\beta_1, \beta_2, \ldots, \beta_m] & \text{if } t \geqq m \end{cases}.$$

A5: for $0 \leqq j \leqq m$, $\alpha_{j-1} > \delta\alpha_j$.

Thus, in many cases, we find the optimal accession policy is to take in the smallest number accessions that is consistent with both the lower bound on accessions ($x_t \geqq 0$) and meets requiremen

$$\sum_{j=1}^{t} \alpha_{t-j} x_j \geqq z_t - y_t.$$

In addition, this thrifty policy is shortsighted; it does not require knowledge of net requiremen $z_{t+1} - y_{t+1}$ etc. in future periods (e.g., $t+1, t+2$) in order to calculate the optimal x_t.

APPROXIMATELY OPTIMAL ACCESSION SCHEDULES

In some cases, it is not possible to obtain the optimal policy in the simple form described in Section
'hese situations typically involve a large legacy and sharp reductions in requirements in the first
periods. In other cases, difficulties arise when the survivor fractions measure a contribution to
ctiveness and are not necessarily decreasing with time. The approximation procedure outlined in
section is designed to handle these cases by providing a simple and finite linear program that
roximates the optimal solution of the infinite linear program.

This section will briefly describe three methods for calculating approximately optimal solution
he infinite horizon optimization problem. Each of the three methods is based on a partition of
original infinite problem into a T period finite problem followed by an infinite problem that com-
ices at time $T+1$. The hope is that the system will settle down enough so that the problem starting
ime $T+1$ will have a nonnegative equality solution regardless of the choice of $(x_1, x_2, \ldots, x_T)$.
The first method simply ignores the decisions and constraints for time $T+1$ onwards. Thus, we
solving problem (29).

$$\text{Minimize} \sum_{t=1}^{T} cx_t\delta^{t-1}$$

$$\text{Subject to} \sum_{j=1}^{t} \alpha_{t-j}x_j \geqq z_t - y_t \qquad \text{for } 1 \leqq t \leqq T.$$

$$x_t \geqq 0$$

ıough this procedure is quite simple, it can lead to optimal programs that save in periods 1 through
y presenting difficult initial conditions for the second problem that commences at time $T+1$.
ce the problem that starts at time $T+1$ is not explicitly considered in the objective, there is no
alty to deter this type of behavior.

The second method attempts to provide a smooth transition to equilibrium by fixing accessions
heir equilibrium value for periods $T+1$ onward. The assumption is that $z_T = z_t$ for $t \geqslant T$, and that

$$x_t = z_T \sum_{j=0}^{m} \alpha_j \qquad \text{for all } t \geqslant T.$$

s, the accessions in periods $1-m, 2-m, \ldots, -1, 0$ and $T+1, T+2, \ldots$, are all known. We
st determine the accessions in periods 1 through T in order to satisfy the lower bound requirement
he first $T+m$ time periods. This leads to a linear program with $T+m$ inequality constraints and
onnegative variables $x_1, \ldots, x_T$. The dual linear program, with T inequality constraints and
m nonnegative variables, is easier to solve. Unfortunately, this truncation procedure has not
n effective in numerical examples we have solved to date. We frequently obtain relatively low values
x_{T-1}, x_{T-2}, etc. and a relatively large value of x_T. In effect, the program satisfies the boundary re-
ction by making a last period correction. This behavior is contrary to the smooth transition to
ilibrium that the model was designed to produce.

The third method is based on the theory developed previously on optimality conditions. The lower

bound on the optimal value of the infinite horizon problem that starts at time $T+1$ must consi the additional legacy due to the accessions in periods 1 through T. The bound is

$$\sum_{j=T+1}^{\infty} c\mu\delta^{j-1}\left(z_j - y_j - \sum_{k=1}^{T} \alpha_{j-k}x_k\right).$$

If we add the value of the first T periods, $\sum_{k=1}^{T} c\delta^{k-1}x_k$, and rearrange terms, we obtain a lower bou for the original infinite problem in terms of the decision variables $x_1, \ldots, x_T$.

$$\sum_{k=1}^{T} c\left(\delta^{k-1} - \delta^T\mu \sum_{l=1}^{m+k-T} \delta^{l-1}\alpha_{l+T-k}\right)x_k - c\sum_{j=T+1}^{\infty} \mu\delta^{j-1}(z_j - y_j).$$

Notice that the expression on the right is a constant, independent of the decisions in periods 1 throu T. Thus, the linear program we solve is

$$\text{Minimize} \sum_{k=1}^{T} c\left(\delta^{k-1} - \delta^T\mu \sum_{l=1}^{m+k-T} \delta^{l-1}\alpha_{l+T-k}\right)x_k$$

(30)

$$\text{Subject to} \sum_{k=1}^{t} \alpha_{t-k}x_k \geq z_t - y_t \qquad t = 1, 2, \ldots, T.$$

$$x_t \geq 0.$$

Notice that the third method is similar to the first, except for the objective which explicitly conta a penalty cost on the legacy created for the infinite problem starting at time $T+1$.

In our calculations, we have used the third method. There are two main reasons for this: fi the lower bound is exact when the equality solution is optimal for the problem commencing in per $T+1$; second, the calculations we have made using actual survivor fractions have led to realis answers.

The dual linear program for (30) is stated below.

$$\text{Maximize} \sum_{t=1}^{T} u_t(z_t - y_t)$$

$$\text{Subject to} \sum_{j=0}^{T-t} \alpha_j u_{t+j} + v_t = \left(\delta^{t-1} - \delta^T\mu \sum_{l=0}^{m+t-T} \delta^{l-1}\alpha_{l+T-t}\right)c$$

$$u_t \geq 0, \qquad v_t \geq 0 \qquad \text{for } t = 1, 2, \ldots, T.$$

The interpretation of the dual variables is straightforward: u_t is the marginal change in the optir cost of increasing net requirements $(z_t - y_t)$, and v_t is the marginal change in the optimal cost associat

increasing the lower bound on accessions in period t. The dual constraints can be interpreted as quation

$$\sum_{j=0}^{m} \alpha_j u_{t+j} + v_t = c\delta^{t-1}$$

e u_t and v_t for $t \geq T$ have been assigned the values $c\delta^{t-1}\mu$ and 0, respectively.

The example below was solved for the rating ET, electronics technician, using the survivor fraction age distribution data found in Table 2 of [3]. We assumed a lower bound of 1,750 accessions per $5c=1$, a discount factor $\delta=0.95$, and $z_t=16{,}000$ for $t \geq 6$.

Since z_t is constant for $t \geq 6$ and the survivor fractions α_j are nonincreasing for the ET rating, now that the equality solution is optimal for the problem beginning in period 6 regardless of the ssions in periods 1 through 5. Thus, the solution above is optimal for the infinite horizon problem.

	Time 1	2	3	4	5	
z_t	20,000	18,000	16,000	16,000	16,000	Manpower Requirements
x_t	2,112	1,750	1,750	1,750	2,828	Accessions
$\sum_{j=0}^{m} \alpha_j x_{t-j}$	20,000	18,363	16,922	16,000	16,000	Actual Manpower Inventory
u_t	0.5	0.0	0.0	0.19	0.18	Marginal Cost of Requirement
v_t	0.0	0.35	0.2	0.0	0.0	Marginal Cost of Accessions Lower Bound

APPENDIX

PROOF OF THEOREM 1: The proof rests on two observations on the T period approximation, elow, of the infinite problem.

$$\text{Minimize} \sum_{t=1}^{T} cx_t\delta^{t-1}$$

$$\text{Subject to} \sum_{j=1}^{t} \alpha_{t-j}x_j \geqq z_t - y_t \qquad \text{for } 1 \leqq t \leqq T.$$

$$x_t \geqq 0.$$

T period dual is

$$\text{Maximize} \sum_{t=1}^{T} u_t(z_t - y_t)$$

$$\text{Subject to} \sum_{j=0}^{T-t} u_{t+j}\alpha_j \leqq c\delta^{t-1} \qquad \text{for } 1 \leqq t \leqq T$$

$$u_t \geqq 0$$

e we interpret $\alpha_j=0$ if $j > m$.

Note first that any feasible solution x_t of the infinite problem has x_t for $1 \leqq t \leqq T$ feasible for In addition, since $\alpha_j \geqq 0$, any u_t that is feasible for the infinite problem will have u_t for $1 \leqq t$ feasible for (2). As a second point, note that any feasible solution of (2) must have $u_t \leqq c\delta^{t-1}/\alpha_0$. T the sum

$$\sum_{t=1}^{\infty} u_t(z_t - y_t)$$

converges for any feasible solution to the infinite problem. From the weak duality theorem for prob (1) and (2), we have

$$\sum_{t=1}^{T} u_t(z_t - y_t) \leqq \sum_{t=1}^{T} cx_t\delta^{t-1}.$$

From our comments above, it follows that this relation must hold for the infinite problem; i.e., for e feasible solution x_t and u_t of the infinite primal and dual, we must have

$$\sum_{t=1}^{\infty} u_t(z_t - y_t) \leqq \sum_{t=1}^{\infty} cx_t\delta^{t-1}. \tag{3}$$

If one can find infinite solutions x_t and u_t such that (3) is satisfied as an equality, then they mus optimal solutions to the infinite primal and dual. We now demonstrate that feasible primal and solutions to the infinite problem have equal objective values if and only if the complementary slack conditions described in (3:9) hold. For all T and infinite solutions x_t, u_t, we have

$$\sum_{t=1}^{T} u_t(z_t - y_t) \leqq \sum_{t=1}^{T} u_t\left(\sum_{j=1}^{t} \alpha_{t-j}x_j\right) \leqq \sum_{t=1}^{T}\left(\sum_{j=0}^{m} u_{t+j}\alpha_j\right)x_t \leqq \sum_{t=1}^{T} cx_t\delta^{t-1}. \tag{4}$$

As $T \to \infty$, the term on the left converges. If complementary slackness holds, then

$$\infty > \sum_{t=1}^{\infty} u_t(z_t - y_t) = \sum_{t=1}^{\infty} u_t \sum_{j=1}^{t} \alpha_{t-j}x_j = \sum_{t=1}^{\infty}\left(\sum_{j=0}^{m} u_{t+j}\alpha_j\right)x_t = \sum_{t=1}^{\infty} cx_t\delta^{t-1}.$$

The first and third equality follow from the complementary slackness conditions. The second equ follows from the fact that

$$\sum_{t=1}^{\infty} u_t \sum_{j=1}^{t} \alpha_{t-j}x_j$$

is finite and nonnegative. If complementary slackness does not hold, i.e., x_t and u_t do not satisfy then for some T the first or third relation in (4) will be a strict inequality, and it is apparent the solu could not have equal objective value.

To this point, we have shown that the optimal value of the primal infinite problem is greater or equal to the optimal value of the dual infinite problem and that equality of value occurs if and

ıplementary slackness conditions are satisfied. The proof is completed by demonstrating that the e problems have optimal solutions with equal values.

)efine x_t^T and u_t^T as optimal solutions of (1) and (2), and complete these sequences by setting), $u_t^T = 0$ for $t > T$. The sequence u_1^T is bounded; thus there exists a point u_1 and a subsequence the positive integers such that $u_1^T \to u_1$ for T in S_1. In general, the sequence u_t^T for T in S_{t-1} is ded; therefore, there exists a subsequence S_t of S_{t-1} and a point u_t such that $u_t^T \to u_t$ for T in S_t. is manner, we can construct an entire sequence u_t. We now demonstrate that there exists a nce of integers R_0 such that

$$\sum_{t=1}^{\infty} \left| u_t^T - u_t \right| \to 0$$

in R_0. Let n be arbitrary and choose $k > n$ such that $n\delta^k < \alpha_0(1-\delta)/c$. This will imply that

$$\sum_{t=k+1}^{\infty} \left| u_t^T - u_t \right| < 1/2n$$

T. Now choose T_n in S_k such that $T_n > T_{n-1}$ and

$$\sum_{t=1}^{k} \left| u_t^T - u_t \right| < 1/2n.$$

onclusion follows for the sequence T_n.

or each T, we have

$$c\sum_{t=1}^{T} \delta^{t-1} x_t^T = \sum_{t=1}^{T} u_t(z_t - y_t) \leqq c/\alpha_0 \sum_{t=1}^{\infty} \delta^{t-1} \mid z_t - y_t \mid .$$

that $u_t^T = 0$ if $z_t - y_t < 0$. Therefore, the sequences x_t^T are bounded above for all t. Let R_1 be a quence of R_0 and x_1 be a point such that $x_1^T \to x_1$ for T in R_1. As before, let R_t be a subsequence of nd x_t be a point such that $x_t^T \to x_t$ for T in R_t. Let t be arbitrary, and consider $T > t$. The finite lementary slackness conditions below must hold.

$$u_t^T \left[\sum_{j=1}^{t} \alpha_{t-j} x_t^T - (z_t - y_t) \right] = 0,$$

$$\left[\sum_{j=0}^{T-t} u_{t+j}^T \alpha_j - c\delta^{t-1} \right] x_t^T = 0.$$

consider the limit of these relations as T in R_t approaches infinity. We see that the infinite ementary slackness conditions hold. In addition, it is easy to demonstrate that x_t and u_t are

feasible solutions of the infinite horizon primal and dual. Therefore, we have shown that optin solutions of the infinite problem exist with equal objective value.

Q.E.

REFERENCES

[1] Duffin, R. J. and L. A. Karlovitz, "An Infinite Linear Program with a Duality Gap," Manageme Science, Vol. *12* (1965).

[2] Feller, W., *Introduction to Probability Theory and Its Applications* (1968), Vol. *1*, Chap. *11*, 3 Edit.

[3] Grinold, R. C. and D. S. P. Hopkins, "Duality Overlap in Infinite Linear Programs," Journal Mathematical Analysis and Applications, Vol. *41*, No. *2* (Feb. 1973).

[4] Grinold, R. C. and R. M. Oliver, "An Interactive Manpower Planning Model," ORC *73–22*, O erations Research Center, University of California, Berkeley (1973).

THE ASSIGNMENT AND SEQUENCING OF OPERATIONS ON A CREW-SERVED PROJECT

R. L. Bulfin
University of Arizona
and
R. G. Parker
Georgia Institute of Technology

ABSTRACT

This paper is concerned with assigning and sequencing a set of activities for some or all members of a crew of operators so that the completion time of all such operations is minimized. It is assumed that each of the operators in the crew possesses, initially, certain tasks that only he can perform. A branch-and-bound scheme is proposed to treat the problem, and suitable computational experience is provided.

RODUCTION

We consider here a problem which involves a crew or group of operators which perform tasks that rise some overall project or job. Typical of such a system would be the operators on a crew-served ons system or perhaps the crew of a space flight mission. With such a system it would be likely each member would possess responsibility for certain tasks. More specifically, each member l have certain tasks that only he could perform. In addition, there would be other tasks that any operators could perform. Assuming that it is desirable to perform all operations of the project nimum time where the operations are considered sequence and operator dependent, the problem, y stated, is one of assigning operations to each operator and sequencing the resulting assigned tions for each such that the maximum completion time over all operators is minimized. In the nt paper, precedence constraints are not considered.

While the description above draws upon the idea of a crew of people, the problem at hand might e a group of machines, each of which are partially loaded with jobs and to which other jobs are allocated or assigned. It is of interest to note that for such a problem, referred to as a parallel-ssor problem [2], it is easy to construct an integer program where it is assumed that all processing are sequence independent. Let a project consist of a set of operations, β. Further assume that are K crew members (machines) and that each member, k, must perform some set of operations, uch that $\beta_k \subset \beta$. Let $|\bigcup_k \beta_k| < |\beta|$ such that there are some operations which are free in the that they can be assigned to any crew member. Denote such a set by $\hat{\beta}$ where $\hat{\beta} = \beta - \{\bigcup_k \beta_k\}$.

The problem can be formulated as follows:

$$\min Z$$

$$\text{subject to } Z \geq \sum_{i \epsilon \beta_k} \tau_{ik} + \sum_{i \epsilon \beta_k} \tau_{ik} x_{ik}; \qquad k = 1, 2, \ldots, K$$

$$\sum_{k} x_{ik} = 1; \qquad i \epsilon \hat{\beta}$$

$$x_{ik} \epsilon \{0, 1\}; \qquad i \epsilon \hat{\beta}, \qquad k = 1, 2, \ldots, K,$$

where x_{ik} is 1 if task i is assigned to crew member k and 0 otherwise. Note that τ_{ij} is the processi time of operation i by crew member j.

The formulation above is similar to that given in [2] for the parallel-processor (machine) proble Although existing computational experience is sparse for such a formulation, it can be hypothesiz that the combinatorial property surrounding the problem would force an increase in the integer progra not concomitant with the physical situation for which it is a model. Moreover, while the problem interest here involves sequence dependencies, the above formulation is not sufficient, and a computational difficulty surrounding the integer program would be compounded if the "sequencin aspect were included. Hence, we consider an alternate approach to the problem.

Consider new sets β_k' which possess the tasks from β_k augmented by those from $\hat{\beta}$ which are a assigned to operator k. Note $\beta_k \subseteq \beta_k'$. Also, let us denote the sequencing problem for the operatio in a given β_k' by P_k'. Recall that it is desired that the operations be ordered for each crew member. addition, assume that all processing times are given in a matrix T. Finally, define a set α which contai tasks or operations that are unassigned. When $\alpha = \phi$, all tasks have been assigned and a feasib solution results.

At this point in the development, it should be pointed out that the inherent sequencing proble that must be solved when all sets β_k' have been constructed can be perceived as single machine sched ing problems with sequence dependent processing times. In such a case, let us assume that there some initial "startup" and final "shutdown" times before the first and final operations (for each β respectively. We can then, without loss of generality, add a dummy operation to each β_k' and solve wh then evolves as a traveling salesman problem (for each β_k'). The solution of a traveling salesm problem is a "tour," or a closed cycle, of all operations, including the dummy. Assuming then that o starts with the dummy, traverses the cycle and returns to the dummy, it is evident that the objecti of ordering the operations subject to initial startup and final shutdown is accomplished.

Suppose the solution to a given problem P_k' has value Z_k'. Then a lower bound on the completi time for crew member k, given that an additional task i is assigned to k, can be given by $B(i; k)$ whe

$$B(i; k) = Z_k' - L_k' + \min_{j \epsilon \alpha \cup \beta_k^\circ} \tau_{ij} + \min_{j \epsilon \alpha \cup \beta_k^\circ} \tau_{ji},$$

where L_k' is the largest τ_{ij} over all ordered pairs (i, j) in the sequence of operations for operator The bounds $B(i; k)$ can be used to determine candidate problems which arise from the selection

task i which is fixed (removed from α) for some crew member k. For every such assignment , another candidate problem arises in $(\overline{i;\ k})$, which denotes the prohibition of the assignment of to member k. A bound on $(\overline{i;\ k})$ can be given by $B(\overline{i;\ k})$ such that

$$B(\overline{i;\ k}) = \min_{\substack{j=1,\,2,\,\ldots,\,K \\ j\neq k}} B(i;\ j).$$

e, the assignment chosen over all candidates $(i;\ k)$ is the one which exhibits the highest bound on . It is worth noting that once an assignment has been chosen and a candidate problem identified, dated lower bound on the completion time of all tasks can be obtained from the new values of k which result from the new problem created by updating specific β'_k.

he computation of lower bounds on free assignments allows for a general branch-and-bound ne whereby candidate problems are identified as described above and stored in a list θ for further it should it be apparent that such a candidate problem possesses some prospect of leading to a on superior to some incumbent. Following, the notions of this procedure are formalized in explicit thmic terms, after which a small sample problem is considered in order to demonstrate its cation.

PUTATIONAL ALGORITHM

he concept of a branch-and-bound procedure described above can be specified by the following y-step computational procedure.

TEP 0. Initialization

0.1 Set $\beta'_k = \beta_k \cup \{x\}$; $k = 1, 2, \ldots, K$. Operation x is a dummy operation.

0.2 Solve P'_k ; $= 1, 2, \ldots, K$. Denote the solution to P'_k; $\forall k$ by Z'_k. Let the maximum link in each sequence be L'_k. Let $\alpha = \hat{\beta}$ and $Z^+ = \infty$. Go to Step 5.

TEP 1. Select a candidate problem

1.1 If $\theta = \phi$, stop. The incumbent is optimal.

1.2 Let θ_n be the member of the list θ which exhibits the lowest bound.

1.3 If θ_n prohibits an assignment, go to Sept 5.

1.4 Update α such that

$$\alpha \leftarrow \alpha - \{i\}, \qquad \text{such that } (i;\ k) = \theta_n.$$

TEP 2. Solve a sequencing problem

2.1 Let task i be assigned to processor k by the current candidate problem.

2.2 Augment β' such that $\beta'_k \leftarrow \beta'_k \cup \{i\}$.

2.3 Solve P'_k and denote its solution value by Z'_k. Let the maximum link in the sequence be L'_k.

TEP 3. Attempt to fathom θ_n

3.1 If $Z'_k < Z^+$, go to Step 4.

3.2 Delete θ_n from θ and go to Step 1.

TEP 4. Check for feasibility

4.1 If $\alpha = \phi$, go to Step 6.

STEP 5. Separate the candidate problem

5.1 Compute bounds $B(i; k)$ for each free assignment.

5.2 Choose a separating assignment $(i; k)$.

5.3 Determine bounds for both new candidate problems and add them to θ if necessary. (
to Step 1.

STEP 6. Replace the incumbent solution

6.1 Let

$$Z^+ = \max_{k=1, 2, \ldots, K} [Z'_k]$$

$$\beta_k^+ = \beta'_k; \qquad k = 1, 2, \ldots, K.$$

6.2 Delete all candidate problems from θ with bounds no better than Z^+ and go to Step 1.

SAMPLE PROBLEM

Consider a small problem involving three crew members and a total of eight tasks. Specifical
consider the following: $\alpha = \{1, 2, 3\}$, $\beta_1 = \{4, 5\}$, $\beta_2 = \{6\}$ and $\beta_3 = \{7, 8\}$. Accordingly, crewmemb
one must perform operations four and five, member two must perform operation six, and memb

TABLE 1. *Data for Sample Problem*

	1	2	3	4	5	6	7	8	9
1	∞	9	5	9	8	6	8	3	5
2	3	∞	4	3	6	8	3	6	4
3	5	4	∞	6	6	7	4	5	2
4	4	5	5	∞	3	∞	∞	∞	4
5	3	2	4	5	∞	∞	∞	∞	3
6	3	3	5	∞	∞	∞	∞	∞	4
7	3	5	3	∞	∞	∞	∞	6	6
8	3	6	2	∞	∞	∞	5	∞	3
9	5	5	2	5	8	9	4	7	∞

three operations seven and eight. Alternately, operations one, two, and three can be performed by a
of the three crew members. The processing times for all operations can be given by matrix T in Table

STEP 0. The procedure is initialized by assuming each set β'_k to be the original sets β_k augment
by the dummy operation, say operation nine. The problems P'_1, P'_2, and P'_3 yield solutions [9–4–5–9
[9–6–9], and [9–7–8–9] with values $Z'_1 = 11$, $Z'_2 = 13$, and $Z'_3 = 13$ respectively. The longest link
each sequence, L'_1, L'_2, and L'_3, is 5. 9. and 6. The value of the incumbent is set initially at ∞. Note th
$\alpha = \{1, 2, 3\} = \hat{\beta}$.

STEP 5. Bounds are computed for each free assignment such that $B(1; 1)$, for example, aris
such that

$$B(1; 1) = Z'_1 - L'_1 + \min_{j \in \alpha \cup \beta'_1} \tau_{1j} + \min_{j \in \alpha \cup \beta'_1} \tau_{j1},$$

e $B(1;\ 1)=11-5+5+3=14$. The other values are computed in similar fashion and can be as follows:

$$\begin{array}{lll} B(1;\ 2)=12 & B(2;\ 1)=11 & B(3;\ 1)=10 \\ B(1;\ 3)=13 & B(2;\ 2)=10 & B(3;\ 2)=8 \\ & B(2;\ 3)=14 & B(3;\ 3)=11. \end{array}$$

Assignment (1; 2) is chosen such that $(\overline{1;\ 2})=13>(\overline{i;\ k})$; $\forall\,(i;\ k)\neq(1;\ 2)$. Hence, two candidate ems are created: $\theta_1=(1;\ 2)$ and $\theta_2=(\overline{1;\ 2})$.

STEP 1. Since the candidate list is not empty, θ_1 is selected for exploration. Note that θ_1 is selected θ_2 since the latter prohibits an assignment. The set α is updated such that $\alpha=\{2, 3\}$.

STEP 2. Observe that β_2' is augmented such that $\beta_2'=\{1, 6, 9\}$. The new problem P_2' is solved ng the solution [9–1–6–9] with value $Z_2'=15$.

STEPS 3, 4, 5. An attempt to fathom $(Z_2'<Z^+)$, the candidate problem, indicates the computation e following bounds:

$$\begin{array}{ll} B(2;\ 1)=11 & B(3;\ 1)=10 \\ B(2;\ 2)=15 & B(3;\ 2)=13 \\ B(2;\ 3)=14 & B(3;\ 3)=11. \end{array}$$

paration is made, creating candidate problems θ_1 and θ_3 such that $\theta_1=\{(1;\ 2)\,(2;\ 1)\}$ and $\theta_3=\{(1;$ $\overline{1})\}$. Since both problems evolve from (1; 2) where $Z_2'=15$, any feasible solution emanating from is bounded by 15.

STEP 1. A new candidate problem θ_2 is chosen.

STEP 5. The candidate problem is separated on (1; 3) and $(\overline{1;\ 3})$ (note that $(\overline{1;\ 3})$ can be replaced 1)). The bound on each member of the separation is 13 and 14 respectively. Hence $\theta_2=\{(\overline{1;\ 2})$ } and $\theta_4=\{(\overline{1;\ 2})(1;\ 1)\}$.

STEP 1. Candidate problem θ_2 is chosen. $\alpha=\{2, 3\}$.

STEPS 2, 3, 4. The set β_3' is updated to $\{1, 7, 8, 9\}$, and problem P_3' is solved such that sequence -1–8–9] arises with $Z_3'=13$. Since $Z_3'<Z^+$ and $\alpha\neq\phi$, the procedure moves to Step 5 for separa- The entire process continues in similar manner and is summarized by the tree in Figure 1. The on can be given by sets β_k^+ such that $\beta_1^+=\{2, 4, 5, 9\}$, $\beta_2^+=\{3, 6, 9\}$ and $\beta_3^+=\{1, 7, 8, 9\}$ with 14, $Z'=13$ and $Z'=13$ where $Z^+=\max[14, 13, 13]=14$. The final sequences which yield the s Z_k' are as follows:

[9–4–5–2–9],
[9–3–6–9],
[9–7–1–8–9].

of interest to note that there may be alternative sequences yielding similar values, e.g. [9–2– 9].

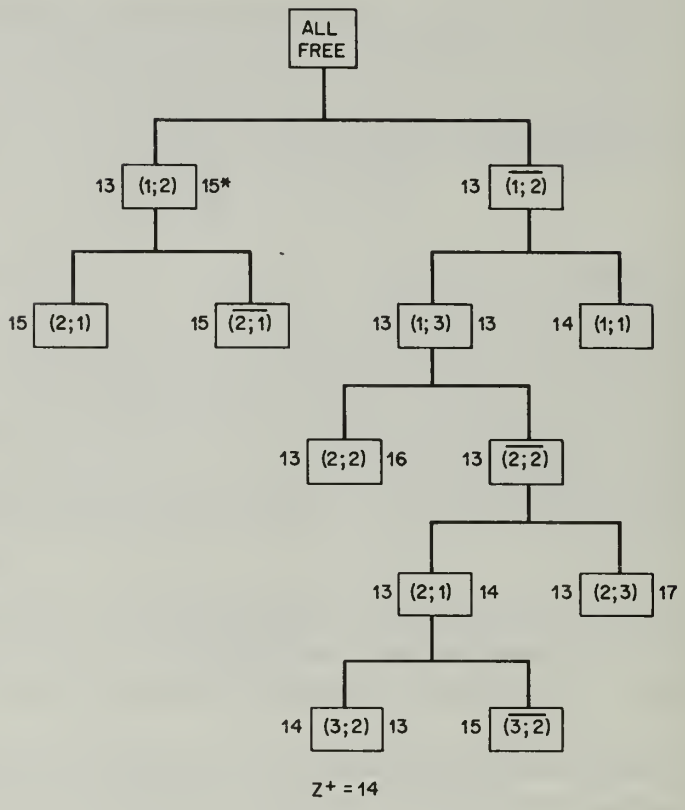


FIGURE 1. Solution tree for sample problem.

COMPUTATIONAL EXPERIENCE

The algorithm was coded in FORTRAN and several problems were run on a U1108 facility. A problems were generated so that the triangularity property held. Such a convention was adopted order that the addition of a task to a sequence would not result in a lower completion time. It shou also be noted that all sequencing problems were solved using a procedure similar to Little, et al. [5 The procedure is, fundamentally, a branch-and-bound technique without backtracking [1]. Hence, th solution to the sequencing problems cannot be claimed to be optimal. Of course, at the expense additional computational effort, a backtrack could be implemented in [1] without difficulty. All comput tional experience is summarized in Table 2. It is important to note that because the fundamental pr cedure used to solve the sequencing problems P_k' is generally well known, we do not include an explic specification of its application in this paper. In this same light, care is taken to point out that oth general procedures could be substituted in place of that in [1]. However, although the computation experience may be altered somewhat, the nature and construction of the entire procedure remai unaffected.

A slight modification to the algorithm as presented above was made. It was desired to obtain a initial solution in hopes of reducing the storage requirements of the branch-and-bound procedure. Th was done by using a LIFO choice of the candidate problems to be investigated. When a separatio

TABLE 2. *Summary of Computational Experience*

Problem number	Size*	Solution initial/final	Quality	Time** initial/final
1	2/2/2	72/70	0.97	0.174/0.201
2	2/6/2	118/118	1.00	0.148/0.154
3	2/10/10	188/184	0.98	0.369/0.738
4	5/2/2	84/82	0.98	0.145/0.238
5	5/6/2	142/114	0.80	0.208/0.315
6	5/10/10	160/158	0.99	0.714/2.468
7	10/2/2	134/132	0.99	0.314/5.190
8	10/6/2	148/148	1.00	0.437/13.397
9	10/10/2	158/158	1.00	0.645/.650
10	10/10/10	194/186	0.96	1.682/6.072
11	15/2/2	142/114	0.80	0.583/13.662
***12	15/6/2	166/—	—	0.873/—
13	2/2/2/2	88/88	1.00	0.147/0.153
14	2/10/10/10	178/174	0.98	0.549/0.963
15	2/6/2/2	136/136	1.00	0.139/0.145
16	5/2/2/2	104/104	1.00	0.149/0.366
17	5/6/2/2	110/110	1.00	0.168/0.173
18	5/10/10/10	192/184	0.96	0.870/1.370
19	10/2/2/2	112/94	0.84	0.362/2.183
20	10/10/2/2	164/164	1.00	0.380/0.385
21	10/10/10/10	204/196	0.96	1.622/71.035
22	15/2/2/2	118/98	0.83	0.577/58.845
23	15/10/2/2	148/140	0.94	0.653/1.980
24	2/2/2/2	90/90	1.00	0.132/0.138
25	2/10/10/10/10	172/172	1.00	0.472/0.477
26	5/2/2/2/2	94/90	0.96	0.211/0.290
27	5/5/2/2/2	104/98	0.94	0.272/1.916
28	5/10/10/10/10	176/176	1.00	0.873/0.878
***29	10/2/2/2/2	104/—	—	0.279/—
30	10/5/2/2/2	108/106	0.98	0.293/18.560
31	10/10/10/10/10	190/178	0.94	1.291/31.281
***32	15/2/2/2/2	116/—	—	0.470/—
33	15/7/2/2/2	128/124	0.97	0.643/27.565
34	2/2/2/2/2/2	88/88	1.00	0.137/0.158
35	5/2/2/2/2/2	94/94	1.00	0.173/0.178
36	5/10/10/10/10/10	160/160	1.00	0.870/0.874
***37	10/2/2/2/2/2	110/—	—	0.282/—
38	10/10/10/10/10/10	168/158	0.94	1.455/23.768
***39	15/2/2/2/2/2	114/—	—	0.492/—
40	15/7/2/2/2/2	112/112	1.00	0.680/0.686

*Problem size given by sets $\hat{\beta} \mid \beta_1 \mid \beta_2 \mid \ldots \mid \beta_K$.
**Time is in *CPU* seconds.
***Storage exceeded.

able was determined (by the stated procedure), the resulting candidate problem that made an gnment $(i; k)$ was chosen as the current candidate problem. In this way, a feasible solution was ined very quickly. After a feasible solution was obtained, the algorithm reverted to choosing the didate problem with the lowest bound.

A primary computational shortcoming of the algorithm appears to be in the excessive number of es required for exploration. Provision was made to store 5,000 nodes, yet, as exemplified by problems

12. 29. 32. 37. and 39 (see Table 2), such a space allocation may not be sufficient. A symptom modification would be to increase space or to use bit packing; however, at the heart of the issue wo be the construction of tighter bounds in order to prune the solution tree and reduce the storage of noc

It can also be seen in Table 2 that solution times possess a high variance. It is clear that minimally loaded crewmember can dominate the solution since, naturally, many if not all tasks $\hat{\beta}$ will be assigned to that member. On the other hand, problems where all or several crewmemb are of nearly equal initial loading require substantially greater amounts of computational effort.

In most problems, the first solution was obtained rather quickly. Moreover, the average ratio the first solution value to that at termination was 0.96. Note, however, that included in the statis were only those problems that did not exceed prespecified core or time limitations.

SUMMARY

This paper has been concerned with a problem involving a set of facilities or operators and the t of assignment to and sequencing over the facilities a set of operations such that the completion ti of all operations is made minimal or near minimal. It is assumed that the operators are initially parti loaded with work. Such a problem would seem to have appeal in systems which are serviced by cre However, there would be no reason to exclude from consideration analogous systems such as conventional multiprocessor machine shop problem and the delivery problem [7].

The attack upon the problem has been one of branch-and-bound. However, it may be of inter to investigate the feasibility of an integer-programming type of approach to the problem similar that discussed briefly at the outset. In addition, recent work with a combinatorial algorithm develop to treat a delivery-type problem [4] is being considered in order that it be made applicable to the curr problem. Nothing definitive can be reported at this time, however. Nevertheless, the branch-a bound scheme appears viable in its presented context, especially when certain problem compositio are encountered. It is clear, on the other hand, that there are problems that are not handled w In particular, when parameters τ_{ij} are very close in value, the efficiency of the algorithm is reduc An important reason for such a drawback is the sometimes indiscriminatory nature of the bounds, t prime result of which is the rapid "horizontal" growth of the decision tree. It would seem then that t pursuit of tighter bounds would be time well spent. Also, when the cardinality of $\hat{\beta}$ is of the magnitu or greater than that of sets β_k and when the latter are nearly equal, the computational effort increas substantially. Again, the parameters τ_{ij} can contribute to such a response. Most importantly, howev such problems create many assignments and, hence, numerous potential solutions of equal or nea equal value.

Finally, it may well be desirable to augment the problem considered in the current paper such th operation precedence constraints are included. Such constraints have been defined elsewhere [as being conditional and unconditional in nature where such conditioning is based upon particul operation assignments. In addition, the flexibility of the given system due to the human component m give rise to considerations such as operation preemption or job-splitting. This notion has been co sidered [6] but has not been fully developed. There may, of course, be other system considerations equal or even more criticality. Their pursuit as well as deeper investigation into the problem of th paper would seem to hold merit.

REFERENCES

Ashour, S., J. Vega, and R. G. Parker, "A Heuristic Algorithm for Travelling Salesman Problems," J. of Transportation Research, Vol. *6* (1972).

Baker, K., *Introduction to Sequencing and Scheduling* (John Wiley and Sons, New York, 1974).

Bulfin, R. L., "The Man-Machine Task Allocation Problem with Sequencing Considerations," unpublished Ph. D. Dissertation, Georgia Institute of Technology (1975).

Holmes, R. A., "The Routing and Scheduling of a Class of Postal Vehicles," M.S. Thesis, Georgia Institute of Technology [1975).

Little, J. D. C., K. G. Murty, D. W. Sweeney, and C. Karel, "An Algorithm for the Traveling Salesman Problem," Operations Res., *11*, 972–989 (1963).

Parker, R. G. and T. O. Kvålseth, "The Man-Machine Task Allocation Problem with Sequencing Considerations," Internal Working Paper, Georgia Institute of Technology (1973).

Turner, W. C., P. M. Ghare, and L. R. Fourds, "Transportation Routing Problem—A Survey," AIIE Transactions *6*, 288 (1975).

MULTICHANNEL QUEUEING SYSTEMS WITH HETEROGENEOUS CLASSES OF ARRIVALS*

U. Narayan Bhat

Department of Industrial Engineering and Operations Research
Southern Methodist University
Dallas, Texas

and

Martin J. Fischer

Defense Communications Engineering Center
Reston, Virginia

ABSTRACT

In an integrated telecommunications network, voice and data traffic compete for the same transmission facilities. Assuming Poisson arrivals and exponential service with different rates, analytic expressions are obtained for measures of performance such as blocking probability and average delay under the following operating rule: class 1 traffic behaves as a loss system while class 2 traffic is buffered when all channels are busy. In view of the inordinate amount of computational effort needed when the number of channels is large, simple approximations have been suggested.

NTRODUCTION

Multichannel systems with two classes of customers with different arrival and service rates ommon in practice. A prime example is the system with voice and data traffic sharing the same o of communication channels. The simplest of such systems can be analyzed by assuming Poisson c and exponential service times for both classes. When service rates for these classes are the same, lard techniques of analysis can be used, since there is no need to distinguish between customer s once a customer joins the system. In this case available results for different operating rules have summarized by Fischer [3]. However, when the service rates for the two classes are different, nagnitude of the problem increases tremendously; this is fully illustrated in Kotiah and Slater [6] e a two-server system of this type has been analyzed.

In this paper we consider a multichannel system with two classes of customers, only one of which ffered when all channels are busy. Customers belonging to the other class leave the system with- ervice if they find no free channel at the time of their arrival. As described above, the system is a stic model for data and voice traffic competing for the use of the same transmission facilities. nally data can be buffered for transmission at a convenient time, whereas, if immediate access is vailable, voice traffic withdraws from the system. To be specific, we shall identify the two types of

Research work for this paper was supported by ONR Contract N00014-72A -0296 and the Defense Communications eering Center, Reston, Virginia. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

traffic as follows: class 1—traffic that does not wait, and class 2—traffic that can be buffered. W assume that these two traffic arrivals are in two independent Poisson processes with para λ_1 and λ_2 and their holding times are exponential with means μ_1^{-1} and μ_2^{-1} respectively. These as tions are commonly made in the analysis of telecommunications systems. As far as the voice tr concerned, comparison with actual data was made at an early date (see, Thorndike [8] and [7]). These comparisons still form the basis of the analytical work in this area. With regard to th traffic, no actual comparisons exist, even though the Markovian assumptions are being widel (for instance see, Kleinrock [5]). Nevertheless one can show that system characteristics are se to the variance of the holding time distributions and that exponential distribution having the variance in the Erlangian family of distribution provides a conservative bound for such characte among this class (also see, Fischer [4]). Furthermore, present analysis techniques severely lin ability to derive useable results under non-Markovian assumptions for service when the syste sists of more than one server.

Let the number of channels in the system be s and the buffer size be infinite. Thus class 1 a to a busy system (all channels occupied) leave without receiving service, and those from class queued and serviced on a first-come, first-served basis as channels become free. Arriving custom a nonbusy system are immediately placed on a free channel.

In spite of the Markovian structure and the practical significance of this system [3], the a are not aware of a satisfactory analysis in the generality of the above description. The specia of $s=1$ has been discussed by Cohen [1]. For larger values of s, as seen from the forthcoming ysis, system equations become unmanageable. In view of this, alternate techniques are nee provide at least approximate values for system performance measures. We consider this as our objective in this paper and develop robust procedures for determining approximate values f probability of blocking for class 1 customers and mean delay for class 2 customers. Neverthele exact analysis of the system is also provided. Given enough computer power and computer the exact analysis can be carried through to its end. However, in practice there are severe limit and an approximate analysis becomes quite attractive. Furthermore, we have used the exact an to validate the approximation procedure and establish its robustness.

The next three sections are organized as follows: Section II gives the general solution to channel system and provides exact results for the case $s=2$. Section III outlines the approxim procedure for which numerical examples are given in Section IV.

II. GENERAL SOLUTION TO THE s-CHANNEL SYSTEM

Let P_{ij} ($i=0, 1, 2, \ldots, s; j=0, 1, 2, \ldots$) be the steady state probability that there class 1 customers and j class 2 customers in the system (i.e., either on the channels or in bu Define for $|z| < 1$

$$\pi_i(z) = \sum_{j=0}^{\infty} P_{ij} z^j, \qquad i=0, 1, 2, \ldots, s.$$

For P_{ij}, the following balance of state equations may be written in the usual mann $\lambda_2 < s\mu_2$, then for $i=0, 1, \ldots, s-1; j=0, 1, \ldots, s-1-i$

$$(\lambda_1+\lambda_2+i\mu_1+j\mu_2)P_{ij}=\lambda_1P_{i-1,j}+\lambda_2P_{i,j-1}+(i+1)\mu_1P_{i+1,j}+(j+1)\mu_2P_{i,j+1}$$

$P_{-1,j}\equiv P_{i,-1}\equiv 0$. For $i=0, 1, \ldots, s$ we have

$$\lambda_2+i\mu_1+(s-i)\mu_2)P_{i,s-i}=\lambda_1P_{i-1,s-i}+\lambda_2P_{i,s-i-1}+(i+1)\mu_1P_{i+1,s-i}+(s-i)\mu_2P_{i,s-i+1}$$

or $j\geqslant s-i+1$

$$(\lambda_2+i\mu_1+(s-i)\mu_2)P_{ij}=\lambda_2P_{i,j-1}+(s-i)\mu_2P_{i,j+1}+(i+1)\mu_1P_{i+1,j}.$$

$=0, 1, 2, \ldots, s$, define

$$D_n(z)=-\lambda_2z^2+[\lambda_2+n\mu_1+(s-n)\mu_2]z-(s-n)\mu_2.$$

Equations (1)–(3) we get the following generating functions:

$$D_0(z)\pi_0(z)=\mu_1z\pi_1(z)-[(\lambda_1-s\mu_2)z+s\mu_2]\sum_{j=0}^{s-1}P_{0j}z^j-\mu_2(z-1)\sum_{j=1}^{s-1}jP_{0j}z^j$$

$$D_n(z)\pi_n(z)=(n+1)\mu_1z\pi_{n+1}(z)-[(\lambda_1-(s-n)\mu_2)z+(s-n)\mu_2]\sum_{j=0}^{s-n-1}P_{nj}z^j$$

$$-\mu_2(z-1)\sum_{j=1}^{s-n-1}jP_{nj}z^j+\lambda_1\sum_{j=0}^{s-n}P_{n-1,j}z^{j+1}\qquad(1\leqslant n\leqslant s-2)$$

$$D_{s-1}(z)\pi_{s-1}(z)=s\mu_1z\pi_s(z)-[(\lambda_1-\mu_2)z+\mu_2]P_{s-1,0}+\lambda_1\sum_{j=0}^{1}P_{s-2,j}z^{j+1}$$

$$D_s(z)\pi_s(z)=\lambda_1zP_{s-1,0}.$$

Equation (5) may be rewritten in the following matrix form:

$$A(z)\Pi(z)=B(z),$$

$$A(z)=\begin{bmatrix}D_0(z) & -\mu_1z & 0 & \ldots & 0 & 0\\ 0 & D_1(z) & -2\mu_1z & 0\ldots & 0 & 0\\ 0 & 0 & . & & & \\ . & . & & . & & \\ . & . & & . & & \\ . & . & & & D_{s-1}(z) & -s\mu_1z\\ 0 & 0 & & & 0 & D_s(z)\end{bmatrix}$$

$$B(z)=\begin{bmatrix} b_0(z) \\ b_1(z) \\ \cdot \\ \cdot \\ \cdot \\ b_s(z) \end{bmatrix}; \qquad \Pi(z)=\begin{bmatrix} \pi_0(z) \\ \pi_1(z) \\ \cdot \\ \cdot \\ \cdot \\ \pi_s(z) \end{bmatrix}$$

and for $n=0, 1, \ldots, s$, we have

$$b_n(z)=\lambda_1 \sum_{j=0}^{s-n} P_{n-1,j} z^{j+1} - \mu_2(z-1) \sum_{j=1}^{s-n-1} jP_{nj} z^j - ((\lambda_1-(s-n)\mu_2)z+(s-n)\mu_2) \sum_{j=0}^{s-n-1} P_{n}$$

with $P_{-1,j} \equiv 0$ for all j and, if the upper index of the summation is less than the lower index in tl pression given for $b_n(z)$, its contribution to $b_n(z)$ is zero.

By Cramer's Rule, for $n=0, 1, \ldots, s$, we have

$$\pi_n(z)=\frac{\det(A_n(z))}{\det(A(z))}, \tag{7}$$

where $A_n(z)$ is $A(z)$ with $B(z)$ replacing the nth column. After simplification, for $n=o, 1, \ldots$ Equation (7) yields

$$\pi_n(z)=\frac{\sum_{j=n}^{s} (\mu_1 z)^{j-n} b_j(z) j!/n! \prod_{k=j+1}^{s} D_k(z)}{\prod_{l=n}^{s} D_l(z)} \tag{8}$$

In the generating function (8) there still remains to determine the $s(s+1)/2$ unknowns; P $i=0, 1, 2, \ldots, s-1$ and $j=0, 1, \ldots, s-1-i$. We need to produce $s(s+1)/2$ indepe equations in these unknowns. From Equation (1), $s(s-1)/2$ equations can be obtained for $1, \ldots, s-1$ and $j=0, 1, \ldots, s-2-i$. Another equation may be obtained by the steady carried load condition (or equivalently by the normalizing condition, $\sum_i \sum_j P_{ij}=1$). In steady we must have ([2]), p. 67):

Expected load carried by any server = Probability that the server is busy,

which can be written as

$$\frac{1}{s}[\rho_2+\rho_1 \sum_{i=0}^{s-1} \sum_{j=0}^{s-1-i} P_{ij}] = \sum_{i=1}^{s-1} \frac{i}{s}[P_{0i}+P_{1,i-1}+\ldots+P_{i0}]+\frac{s}{s}[1-\sum_{i=0}^{s-1} \sum_{j=0}^{s-1-i} P_{ij}], \tag{9}$$

where $\rho_i=\lambda_i/\mu_i$, $i=1, 2$. The final $s-1$ equations are given by the $s-1$ roots of $\det(A(z))$ in $(0, 1)$. For $n=1, 2, \ldots, s-1$ let

$$\zeta_n = \frac{\lambda_2 + n\mu_1 + (s-n)\mu_2 - \sqrt{(\lambda_2 + n\mu_1 + (s-n)\mu_2)^2 - 4(s-n)\lambda_2\mu_2}}{2\lambda_2}$$

ζ_n is the unique root of $D_n(z)$ inside $(0, 1)$. Since $\det(A(z)) = \prod_{n=0}^{s} D_n(z)$, it is clear that ζ_n, $n = 1$, . . ., $s-1$, are the unique roots of $\det(A(z))$ inside $(0, 1)$. Equating the zeros of the numerator and minator, we have from Equations (7) and (8) that for $n = 1, 2, \ldots, s-1$,

$$\sum_{j=n}^{s} (\mu_1\zeta_n)^{j-n} b_j(\zeta_n) \frac{j!}{n!} \left\{ \prod_{k=n+1}^{j} D_k(\zeta_n) \right\}^{-1} = 0$$

$\prod_{n+1}^{n} D_k(\zeta_n) \equiv 1.$

The two main measures of performance are the blocking probability, PB, for class 1 and the average in the system, EW, for class 2. It is immediate that

$$PB = 1 - \sum_{i=0}^{s-1} \sum_{j=0}^{s-1-i} P_{ij};$$

ermore, if P_i is the marginal probability of the number of class 1 customers in the system, then from tion (5) we have, for $i = 1, 2, \ldots, s$,

$$P_i = \rho_1 \frac{\sum_{j=0}^{s-i} P_{i-1,j}}{i}$$

$$P_0 = 1 - \sum_{i=1}^{s} P_i.$$

$E(Q_2)$ is the expected number of class 2 customers in the system, then $EW = \lambda_2^{-1} E(Q_2)$, where

$$E(Q_2) = \sum_{n=0}^{s} \pi_n'(1).$$

fferentiating Equation (6) and solving the resulting equations by Cramer's Rule, for $n = 0, 1, \ldots$, finds

$$\pi_n'(1) = \frac{\sum_{j=n}^{s} \mu_1^{j-n} e_j j!/n! \prod_{k=j+1}^{s} D_k'(1)}{2 \prod_{i=n}^{s} D_i'(1)},$$

where

$$e_j = \lambda_1 \sum_{j=0}^{s-n} j(j+1)P_{n-1,j} - 2\mu_2 \sum_{j=0}^{s-n-1} j^2 P_{nj} - \lambda_1 \sum_{j=0}^{s-n-1} j(j-1)P_{nj}$$

$$- 2(\lambda_1 - (s-n)\mu_2) \sum_{j=0}^{s-n-1} jP_{nj} + 2\lambda_2 P_j. \tag{13}$$

The same procedure may be used to determine all the moments of the distribution of the nu of class 2 customers in the system.

As mentioned in Section I, the special case $s=1$ has been solved by Cohen [1]. Specializin results to the case of $s=2$, we have

$$P_{00} + \zeta P_{01} + \left\{ \frac{2}{\rho_2\alpha(1-\zeta)+2} - \left(1 - \frac{\alpha}{\rho_1} + \frac{\alpha}{\rho_1\zeta}\right)\right\} P_{10} = 0, \tag{14}$$

$$(2+\rho_1)P_{00} + (1+\rho_1)P_{01} + (1+\rho_1)P_{10} = 2-\rho_2, \text{ and} \tag{15}$$

$$(\rho_1 + \rho_2\alpha)P_{00} - \alpha P_{01} - P_{10} = 0, \tag{16}$$

where

$$\zeta = \frac{1}{2}\left(\left(1 + \frac{1}{\rho_2\alpha} + \frac{1}{\rho_2}\right) - \sqrt{\left(1 + \frac{1}{\rho_2\alpha} + \frac{1}{\rho_2}\right)^2 - 4/\rho_2}\right). \tag{17}$$

Solving, we get

$$P_{00} = \frac{(\zeta - \eta\alpha)(2-\rho_2)}{\{2+\rho_1 + (1+\rho_1)(\rho_1\rho_2\alpha)\}(\zeta - \eta\alpha) - (1+\rho_1)(1-\alpha)(1+\eta(\rho_1+\rho_2\alpha))}, \tag{18}$$

$$P_{01} = P_{00}\frac{1+\eta(\rho_1+\rho_2\alpha)}{\zeta - \eta\alpha}, \tag{19}$$

$$P_{10} = (\rho_1 + \rho_2\alpha)P_{00} - \alpha P_{01}, \tag{20}$$

with

$$\eta = \frac{2}{\rho_2\alpha(1-\zeta)+2} - 1 + \frac{\alpha}{\rho_1} - \frac{\alpha}{\rho_1\zeta}.$$

These results may be used to give an expression for the probability of blocking $PB = 1 - (P$ $P_{01} + P_{10})$ and the mean delay as in Equation (12). In Section IV we present some numerical res

ıe cases $s = 2$ and $s = 5$, the latter system having been solved by the same direct method described
e.

SOME SIMPLE APPROXIMATIONS

In Section II it was shown that the solution to the s-channel problem is dependent on the solution
set of $s(s+1)/2$ linear equations in $s(s+1)/2$ unknowns. Further, a systematic method of generat-
hese equations was given. Once the equations have been generated, standardized computer
ages could be used for solving these equations. But the problem is that when s gets large, the
ber of equations grows rapidly. For instance, when $s = 5$, the number of equations is 15; when
0, the number of equations is 210. Hence the problem becomes computer limited as the number of
nels increases.

n this light it would be desirable to have some approximations for the blocking probability for
1, P_aB, and the average waiting time for class 2, E_aW, that do not require the solution of large
bers of equations. This section contains some approximations of these system performance meas-
and discusses their properties. Numerical comparisons are made in Section IV.

In section II, for $s = 2$, we showed that the blocking probability for class 1 was a function of α.
ever, analytical expressions and numerical investigations for the blocking probability (see Tables 1,
d 5 of Section IV) indicate that its sensitivity to changes in α is not very pronounced. Using this
basis, we shall use the blocking probability under the assumption $\mu_1 = \mu_2$ (i.e., $\alpha = 1$) as an
ximation for the system in consideration. Writing $\rho = \rho_1 + \rho_2$ and solving the system of equations
ıe case $\mu_1 = \mu_2$, we get the proposed approximation PB as

$$P_aB = \frac{\rho E(\rho, s-1)}{s - \rho_2 + \rho E(\rho, s-1)},$$

e $E(\rho, s)$ is Erlang's loss formula given by

$$E(\rho, s) = \frac{\rho^s/s!}{\sum_{j=0}^{s} \rho^j/j!}.$$

erical considerations in Section IV show that it is indeed a good approximation. However, it should
ted that, if $\alpha < 1 (> 1)$, the approximation P_aB would be an over (under)-estimate of the actual
ing probability.

n order to develop an approximation for the average waiting time, some well known results are
from the standard $M/M/s$ queueing system with one class of arrivals. Letting W_q stand for the
ng time (steady state) in the queue of an arriving customer, and V for the busy period (time interval
g which all s channels are continuously busy), we have

$$E(W_q | W_q > 0) = E(V).$$

[2], pp. 77 and 222 for elaboration of this relationship. By unconditioning we get

$$E(W_q) = E(W)P\{W_q > 0\}, \tag{23}$$

$$= E(V)PB$$

where PB is the probability of finding all channels busy.

In the system under consideration here, if $\mu_1 = \mu_2 = \mu$,

$$E(W) = \frac{1}{s\mu - \lambda_2},$$

and PB is given by (21). Furthermore, it can be shown easily that

$$E(W_q) = \frac{PB}{s\mu - \lambda_2}.$$

Thus for the system considered in this paper, where $\alpha = 1$, the result given by Equation (23) is tru

The following approximation is proposed for the average waiting time for a class 2 custome

$$E_a W = E(V)\, P_a B + \frac{1}{\mu_2}, \tag{24}$$

where $P_a B$ is given by Equation (21).

It remains to determine the expected length of a busy period, $E(V)$, in the system u consideration. Let V_i be the length of a busy period starting with i class 1 customers. By usin standard technique employed in analyzing a busy period [2], a functional relationship satisfied b distribution function of V_i can be developed. However, in the basic system under study, a busy p starts with some class 1 and some class 2 customers. Therefore an approximation is obtaine changing the workload of all class 1 customers present at the first departure point in a busy p into an equivalent class 2 load. This results in the following expression for $E(V_i)$ $i = 0, 1, \ldots,$

$$E(V_i) = \left(\frac{s\mu_2}{s\mu_2 - \lambda_2}\right)\left(\frac{1}{i\mu_1 + (s-i)\mu_2}\right) \tag{25}$$

$$+ (\alpha - 1)\left(i - \frac{i}{i\mu_1 + (s-i)\mu_2}\right)\left(\frac{\lambda_2 + (s-i)\mu_2}{(s\mu_2 - \lambda_2)\ (\lambda_2 + i\mu_1 + (s-i)\mu}\right.$$

In order to determine $E(V)$, we must uncondition the above result by using an initial probab vector. No simple expression for these probabilities could be found, so these probabilities are app mated by a weighted binomial distribution. Thus

$$E(V) = \sum_{i=0}^{s} \binom{s}{i}\theta^i(1-\theta)^{s-i}E(V_i), \tag{26}$$

where $E(V_i)$ is given by (25) and

$$\theta=\frac{\lambda_1(1-P_aB)}{\lambda_1(1-P_aB)+\lambda_2}.$$

For the case where $\alpha=1(\mu_1=\mu_2=\mu)$, $i=0, 1, \ldots, s$,

$$E(V_i)=\frac{1}{s\mu-\lambda_2}$$

$$E(V)=\sum_{i=0}^{s}\binom{s}{i}\theta^i(1-\theta)^{s-i}E(V_i)$$

$$=\frac{1}{s\mu-\lambda_2}.$$

is, the approximation is exact. Furthermore, as the class 1 traffic becomes light ($\lambda_1\to 0$), $\theta\to 0$ and nly contribution to $E(V)$ in (26) is when $i=0$. But from (25)

$$E(V_0)=\frac{1}{s\mu_2-\lambda_2},$$

the approximation is again exact when there is no class 1 load. These properties and some addi- l numerical comparisons are given in Section IV.

NUMERICAL EXAMPLES

In this section some numerical examples are presented using the preceding analysis. Tables 1 2 give the values of the probability of blocking for the cases where $s=2$ and $s=5$, respectively, varying values of α and ρ_2. Tables 3 and 4 are comparisons of the actual average waiting time approximations, as presented in Section III, for $s=2$ and $s=5$, respectively. Again, α and ρ_2 been varied. Finally, Table 5 investigates the sensitivity to α of the blocking probability, average ue length, average waiting time, and average waiting time approximation.

TABLE 1. *Probability of Blocking with*

$s=2, \quad \rho_1=0.5, \quad \mu_2=1$

α \ ρ_2	0.5	1.0	1.5
0.5	0.2480	0.47144	0.72593
1.0	0.25000	0.47368	0.72727
5	0.25743	0.48210	0.73174
10	0.26032	0.48584	0.73347
20	0.26240	0.4890	0.73478

TABLE 2. *Probability of Blocking with*

$s = 5, \quad \rho_1 = 5, \quad \mu_2 = 1$

α \ ρ_2	0.5	1.5	2.5	3.5	4.5
0.5	0.07547	0.20143	0.38421	0.60962	0.86459
1	0.07688	0.206519	0.391016	0.61509	0.866665
5	0.08129	0.22143	0.41067	0.63058	0.87211
10	0.08303	0.22656	0.41736	0.63606	0.873986
20	0.08429	0.23021	0.42214	0.64020	0.87541

TABLE 3.* *Comparison of Waiting Time and Approximation with*

$s = 2, \quad \rho_1 = 0.5, \quad \mu_2 = 1$

α \ ρ_2	0.5	1.0	1.5
0.5	1.124 / 1.101	1.412 / 1.351	2.378 / 2.271
1	1.166 / 1.166	1.473 / 1.473	2.454 / 2.454
5	1.389 / 1.321	1.811 / 1.648	2.932 / 2.649
10	1.587 / 1.363	2.105 / 1.683	3.440 / 2.683
20	1.931 / 1.388	2.5478 / 1.702	4.382 / 2.700

*The actual average waiting time is the upper entry, i.e.

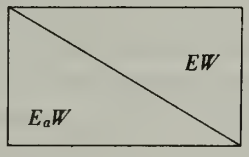
EW
$E_a W$

Tables 1, 2, and 5 show that the probability of blocking is dependent on α, but not to a great e
For instance, in Table 5, as α ranges from 0.0005 to 5000, the probability of blocking change
from 0.2461 to 0.2652. Hence if the approximation suggested in this paper (i.e. $PB = 0.2500$) is
it would be low (high) if $\alpha < 1$ (>1) by at most 0.0152 (0.0039). Analogous comparisons are
available from Tables 1 and 2.

TABLE 4.* *Comparison of Waiting Time and Approximation with* $s=5, \quad \rho_1=5, \quad \mu_2=1$

α \ ρ_2	0.5	2.5	4.5
0.5	1.009 / 1.006	1.105 / 1.044	2.647 / 2.356
1	1.017 / 1.017	1.156 / 1.156	2.733 / 2.733
5	1.062 / 1.137	1.357 / 1.422	3.047 / 3.130
10	1.104 / 1.211	1.482 / 1.477	3.283 / 3.187
20	1.172 / 1.264	1.639 / 1.507	3.664 / 3.217

ıe actual average waiting time is the upper entry, i.e.



'ables 1, 2, and 5 also exhibit another interesting property of the blocking probability with ct to α. As α increases, the blocking probability also increases. This fact becomes more interesting one considers that, when $s=1$, the blocking probability was only a function of ρ_1 and ρ_2; i.e. $(\rho_1+\rho_2)/(1+\rho_1)$. The fact that the blocking probability does not depend on α when $s=1$ e explained as follows. In the one channel case, the class 1 customers may join the system only the server is idle. Since the arrival process of each class is Poisson, the service time of either does not affect the amount of time the system is free to accept class 1 customers. Thus the ng probability for class 1 is independent of α.

Vhen $s>1$, class 1 customers may join the system any time there are $s-1$ or less busy channels. g this time there may be customers from either class 1 or 2 occupying the channels, and so the of their service rates would determine the amount of time the system is available to accept class omers. Hence PB does depend on α when $s>1$.

'ables 3, 4, and 5 also give comparisons of the average waiting time with the approximation. Two of interest with respect to the approximation are noted: first, as $\alpha\to\infty$, the approximation gets , and second, the approximation seems better when $s=5$ than when $s=2$. Numerical investi- s via simulation for the cases where $s>5$ not reported here indicate that as s increases the ıcy of the approximation is about the same as the numerical results given in Table 4 ($s=5$). An

assessment of approximation is left to the discretion of the reader. But certainly one would have agree that both the approximations are quite useful if the main criteria are ease and feasibility computation.

TABLE 5. *Sensitivity to α of System Performance Measures with*

$s = 2, \quad \rho_1 = 5, \quad \rho_2 = 0.5, \quad$ and $\mu_1 = 2$

α	PB	$E(Q_2)$	EW / E_aW
0.0005	0.2461	0.5333	1006.7 / 1000.1
0.005	0.2461	0.5336	106.7 / 100.1
0.05	0.2462	0.5368	10.74 / 10.08
0.5	0.2480	0.5618	1.124 / 1.101
1	0.2500	0.5833	0.5833 / 0.5833
5	0.2574	0.6945	0.1389 / 0.1321
25	0.2629	1.046	0.04187 / 0.02787
100	0.2646	2.212	0.02212 / 0.00705
500	0.2651	8.340	0.01608 / 0.00142
1000	0.2652	15.994	0.01599 / 0.00071
5000	0.2652	77.219	0.01544 / 0.000141

REFERENCES

[1] Cohen, J. W., "Certain Delay Problems for a Full Availability Trunk Group Loaded by Two Tra Sources," Communication News, *16*, 105–113 (1956).

[2] Cooper, R. B., *Introduction to Queueing Theory* (The Macmillan Company, N.Y., 1972).

[3] Fischer, M. J., "A Queueing Analysis of Some Possible Operating Rules for an Integrated T communications Network," DCA System Engineering Facility TN5–72 (Dec. 1972).

[4] Fischer, M. J., "The Waiting Time in the $E_k/M/1$ Queueing System," Operations Res. *22*, 898- (1974).

[5] Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay* (Dover, New York, 19

[6] Kotiah, T. C. T. and N. B. Slater, "On Two-Server Poisson Queues with Two Types of Custome Operations Res. *21*, 597–603 (1973).

[7] Molina, E. C., "Applications of the Theory of Probability to Telephone Trunking Problems," Syst. Tech. J. *6*, 461–494 (1927).

[8] Thorndike, F., "Applications of Poisson Probability Summation," Bell Syst. Tech. J. *5*, 604- (1926).

A MEASURE OF EFFECTIVENESS FOR SENSORS AND STRATEGIES*

Charles E. Antoniak

University of California—Berkeley

ABSTRACT

This paper discusses a possible measure of effectiveness for information sensing equipment, such as radar or sonar, based on its ability to provide information. It is shown that such a measure is particularly appropriate for situations where a sequence of similar limited engagements may occur, as on antisubmarine patrol, for example. In this case the measure expresses the expected rate of gain per engagement of the relative resources of the participants. Rates are calculated for the optimal and certain simpler, but suboptimal, strategies. The measure is illustrated by an exact analysis of a gambling problem and a qualitative treatment of an anti-missile missile allocation problem.

ACKGROUND AND SUMMARY

The operating characteristics of radars and sonars are often given in terms of specification of onic and physical capabilities, such as peak or average power output, pulse repetition frequency, ting bandwidth, and scan rate. These specifications are convenient in the sense that they are iptive of the equipment itself and do not depend strongly on the physical environment or tactical ion in which the equipment will be used. However, they are not a convenient set of descriptors eciding which of several available radars or sonars is "best" to install on a ship or aircraft. More l in this regard would be a description of the ability of the sensor equipment under consideration tect various kinds and sizes of targets, at various distances, under various sea states and orological conditions. The problem would be simple, of course, if one radar or one sonar were rmly better than its competitors under all conditions. This is seldom the case, of course. The more situation is that one radar is superior to another at detecting passive targets, but is more ptible to countermeasures such as jamming or spoofing; or that one sonar achieves longer ranging ilities than another in deep ocean by exploiting thermal ducts, but is virtually useless in a w water, reverberation-limited environment.

What one would like is a single measure of effectiveness that would somehow evaluate the overall lness of some existing or proposed radar or sonar, where a higher rating would infer fairly directly ater tactical utility, at least in some average sense. A type of measure sometimes proposed [5] e based on the equipment's data rate, or ability to provide information. The usual motivation ch a measure is its intuitive and heuristic appeal. It is the purpose of this note to show that some ures of effectiveness based on information rate have more than simply intuitive appeal; they can

This research was supported by the U.S. Navy Electronics Lab, San Diego, Calif., and by the Office of Naval Research Contract N00014-69-A-0200-1051 with the University of California.

be related directly to the advantage to be expected in a tactical situation. The mathematical ba for the proposed measure of effectiveness was originally developed for the analysis of certain gambl problems by Kelly [3], Breiman [1], and Dubins and Savage [2]. These analyses of gambling situati were concerned with determining the optimal betting strategy when a player has acquired ex information about the relative probabilities of the various possible outcomes of play. This paper app some conclusions of the gambling analyses to the allocation of defensive resources in a threat sit tion, i.e., threat evaluation and weapon assignment (TEWA) problems.

A major, surprising conclusion of the gambling analysis cited above is that in some situations, gambler should make bad bets; that is, he should wager some money on possible outcomes wh payoff odds are less than the reciprocal of the probability of that outcome. The kind of gambl situation where this unusual strategy is optimal is one where the gambler expects to play a l sequence of games, not just a single game, and has a continuing supply of information about the relat probabilities of the various possible outcomes. In a military context, the corresponding conclusion that a plan of action which *would* be optimal in a single set-piece engagement would be seriou suboptimal if the engagement were just one of a foreseeable sequence of limited engagements, a that in such situations there may be sound theoretical reasons for "defying the odds."

The gambling model also develops a quantitative measure for the value of extra information i game and relates it to the expected improvement in a gambler's fortunes that would result from using this information and betting an amount of money on each possible outcome according to t optimal allocation prescribed by the theory. The value of information is translated into the expect rate of return it makes possible, and various sources of information can be compared on this bas This paper develops an analogous measure of effectiveness (MOE) for sensors based on the informati they provide and the expected improvement they could produce in a repetitive tactical situation.

Finally, the theory for the gambling problem also makes it possible for the gambler to compu the expected rate of return on a suboptimal strategy which he may want to implement for reasons convenience or practicality. For example, the optimal strategy may call for fractional dollar amou when the house rules require integer valued bets. Similarly, there is a restraint in tactical situatio that ships and aircraft must be dispatched in integer amounts. The MOE developed in this paper e ables a planner to evaluate numerically the various physically realizable strategies that may be ava able to him. In addition, the theory demonstrates that changes in payoff which do not affect the "equi ble" status of the problem and do not change the underlying probabilities, do not require any chan in the allocation of resources. This has implications for the direction of weapon development, inso as it suggests a way of measuring the trade-off between the effectiveness of countermeasures and t probability of their being used.

II. MODEL AND ANALYSIS

This section develops the model and the theory of allocation of resources in a general framewo of possible events and the payoffs that result from assigning resources to some event. In Section I a specific example will be given that illustrates the theory with a simplified model of an anti-miss missile allocation problem.

Let $\mathcal{S}=\{s_1, s_2, \ldots, s_n\}$ be the set of all possible situations, or states of nature, or possible outcom of an experiment. The number n may be large because all the various combinations of elementary si ations must be separately enumerated. Let S be the actual but unknown state of nature, and let p $P(S=s_i)$.

uppose we have a finite amount of resources R, which we can allocate to the various contingencies α_1 to s_1, . . ., α_n to s_n. with the outcome that if s_1 is the true "state of nature," we will end up $a_i\alpha_i$. where a_i is the factor (odds) by which the resources assigned to s_i are multiplied if, in fact, . *All the resources assigned to other states are assumed lost*, as would be the case, for example, if timissile missile were fired at a target which wasn't there.

EFINITION I. A decision problem will be called "fair" or "a fair game" if $a_i = 1/p_i$ for all i.

EFINITION II. A decision problem with payoff odds $\{a_i\}$ is *equitable* if there exists some stent collection of $\{p_i'\}$ such that the odds $\{a_i\}$ are fair with respect to the $\{p_i'\}$.

t is clear that a fair game is equitable and that a decision problem will be "equitable" if $\sum_{i=1}^{n} 1/a_i = 1$, the collection $\{p_i' = 1/a_i\}$ satisfies Definition II. The word "equitable" is used for the situation finition II because, even without knowing the true values of $\{p_i\}$, we could assign $\alpha_i = R/a_i = p_i' R$ ch event s_i. Then, whatever the true value of S, say $S = s_j$, we would end as we started, with $R = R$.

Ve are interested in situations where the original payoff odds a_i are consistent with a fair game probabilities p_i', but subsequent information obtained by the decisionmaker implies that a ent set of probabilities $\{p_i\}$ now holds for the events $\{s_i\}$ even though the original payoff odds are still available. In the gambling models usually considered, this situation could arise if the ler had access to information not available to the odds-makers. He might, for example, have vered a bias in a roulette wheel. In the tactical situation we describe in Section III, the adversary umed to have placed some of his resources at risk in a surprise missile attack which could be ively neutralized by the defender using fewer resources if he knew the direction of the attack.

he mathematical theory developed by Kelly [3] shows that whenever the decision problem is able but the decisionmaker has information that implies that the true probabilities $\{p_i\}$ are ent from $\{1/a_i\}$, then the decisionmaker has an optimal strategy for taking advantage of his nation. Hence, it is important to obtain as accurate information as possible about the true values e p_i's. It is the function of information processors to provide this information. The more mation" a processor provides, the "better" it is. We make this statement more precise as s:

et $\underline{p} = (p_1, p_2, \ldots, p_n)$ be our prior knowledge of the probabilities associated with each which may already be different from $\underline{p}' = (1/a_1, 1/a_2 \ldots)$. Then assume that we acquire onal data through sensors such as radar or sonar. This data is fed into an information processor, output can be characterized as a vector $\underline{p}^* = (p_1^*, \ldots, p_n^*)$. We recognize that for some ssors, such as a maximum likelihood indicator, the output vector may have simply a 1 in the sition and 0's in all other slots. If the processor is a reliable Bayes processor, the $\underline{p}^*$ vector will be the best *a posteriori* estimate of the true state of nature.

lowever, the output of a processor may appear to contain more information than it has. For ple, a binary symmetric channel may output a series of 0's and 1's, but there is a probability of q associated with each symbol. That is, if we receive a 1, we realize that, in fact, the probability 1 was sent is only $1-q$ and that there is probability q that a 0 was sent and an error was made. were to allot our resources as if we were *sure* that a 1 was sent, we would be making a potentially rous mistake. In the more general model given above, we express the unreliability of the informa-

tion processor by saying that for each processor output probability vector $\underline{p}^*$, there corresponds inferred probability vector $\underline{p}^{**}$ which represents the posterior probabilities of the various states nature, given that the processor output was $\underline{p}^*$. In the binary symmetric channel, the relation betw $\underline{p}^*$ and $\underline{p}^{**}$ is relatively simple, e.g. $\underline{p}^*=(1, 0)$ implies $\underline{p}^{**}=(1-q, q)$. In the more general case, t relation may need to be determined empirically. In this paper we assume that in one way or anot the relation between $\underline{p}^*$ and $\underline{p}^{**}$ can be determined.

As an example of the kind of relation that might exist in practice, consider the following simpli model of a Navy Tactical Data System (NTDS) problem. Suppose a destroyer Combat Informat Center (CIC) information processor classifies an attack as coming from in front (s_1), from the starbo beam (s_2), or from the port beam (s_3). Suppose the prior probability $\underline{p}$ of such attacks is $\underline{p}=($(0.25, 0.25), and let the combined operating characteristics of the sensor and intelligence inputs the processor output be summarized in the matrix below.

PROCESSOR I OUTPUT

TRUE STATE \	1	2	3
1	0.8	0.1	0.1
2	0.1	0.8	0.1
3	0.1	0.1	0.8

FIGURE 1

That is, when the attack is actually coming from in front, the processor correctly indicates this percent of the time and incorrectly indicates an attack from the port or starboard beam 10 percen the time each. The situation is similar for the second and third rows of the matrix.

For the prior and matrix given, an output $\underline{p}^* = (1, 0, 0)$ would correspond to an inferred probabi $\underline{p}^{**}=(0.88, 0.\underline{05}, 0.\underline{05})$.† Hence if the processor indicated an attack from the front, there would an 11 percent probability that the attack was in fact coming from the port or starboard beam.

How can we determine whether this processor is to be preferred to one whose operat characteristics are described by the matrix below, for example, which reports frontal attacks m reliably and flank attacks less reliably than the first processor?

PROCESSOR II OUTPUT

TRUE STATE \	1	2	3
1	0.9	0.05	0.05
2	0.1	0.7	0.2
3	0.1	0.2	0.7

FIGURE 2

There are two steps in the derivation of an answer. The first is to define the information provided l processor, and the second is to show a direct relation between the information provided by a proce and the gain in resources a decisionmaker can expect to reap if he uses this information in an opti manner. The fact that the gain is a function only of the information and the practical significance

† For the exact calculations see the Appendix. The notation $0.\underline{05}$ indicates the continued decimal 0.05555 . . . .

n resources, leads us to define a measure of effectiveness of an information processor in terms of formation it provides.

Ve proceed to develop the steps outlined above in reverse order. Suppose the decisionmaker vailable to him an equitable problem with payoff odds $\{a_i\}$ and his current state of knowledge is by $\underline{p}$. He intends to allot a certain fraction α_i of his resources to event s_i, and if $S=s_i$, his total rces after one play are $R_1=a_i\alpha_i R_0$. We assume he must distribute all his resources over the pos- outcomes without reserve. This is not a restriction in an equitable problem because the decision- r can effectively reserve any amount r that he wishes to by allotting r/a_i to every event Si. What- happens, he will recover at least $a_i(r/a_i)=r$. Furthermore, we are considering the case where urrent game is just one in a long sequence, and his object is not only to maximize the final amount resources but also to maximize the probability that his resources do not fall to zero. These con- itions lead us to adopt the optimality criterion of Kelly [3] and evaluate the results of a single play ms of the expected value of the log of the ratio of "resources after" to "resources before." This es us to obtain the expected log of the gain over several plays by adding the expected logs of the over each play. To this end we define the expected gain factor $G(\underline{p}, \underline{a}, \underline{\alpha})$ as a function of the rs of probabilities $\underline{p}$, payoff factors $\underline{a}$, and allocation fractions $\underline{\alpha}$, as

$$G(\underline{p}, \underline{a}, \underline{\alpha}) = E(\log\ (R_1/R_0) \mid \underline{p}, \underline{a}, \underline{\alpha}) = \sum_{i=1}^{n} p(i)\ \log\ [a_i\alpha_i].$$

Ve note that the maximum value of $G(\underline{p}, \underline{a}, \underline{\alpha})$ is obtained by making the allocation $\alpha_i=p_i$, inde- nt of the payoff vector $\underline{a}$. A detailed proof can be found in Kelly [3]; the conclusion follows directly he identity

$$\sum_{i=1}^{n} p_i \log \alpha_i a_i = \sum p_i \log a_i + \sum p_i \log \alpha_i.$$

rst term on the right is independent of $\underline{\alpha}$, and the second term is a familiar entropy expression n to be maximized by $\alpha_i=p_i$. We list below three important consequences of the fact that the al allocation α_i is proportional to p_i, and we shall consider them in turn.

) An information processor will generally change the value of $\underline{p}$, and hence will call for a new tion of resources $\underline{\alpha}$.

2) Changes in the payoff vector $\underline{a}$ that are not accompanied by changes in the probability vector ıot call for changes in resources allocation, even if some of the new odds look attractive.

3) Knowing the optimal allocation and maximum possible rate of gain, we are in a better position luate suboptimal but more easily realizable strategies. In some cases there will be very small nces between optimal strategies and simple suboptimal strategies.

eginning with the first point, we proceed to define a measure of effectiveness for information ssors in terms of the changes in the probability vector $\underline{p}$ that they may produce, with a resulting e in the optimal allocation and its resulting expected gain.

ı this derivation, we assume that the decisionmaker uses the optimal strategy, so that $\alpha_i=p_i$. he maximum expected gain before processing is

$$G_{\max}(\underline{p}, \underline{a}) = \sum_{i=1}^{n} p_i \log p_i a_i$$

$$= \sum p_i \log a_i + \sum p_i \log p_i,$$

and we recognize the second term as $-H(S)$, the negative of the entropy of S. Similarly, when calculate the average expected gain after processing, we obtain

$$E(G_{\max}(\underline{p}, \underline{a})|X) = E(G_{\max}(\underline{p}|\underline{X}, \underline{a})) = \sum_{j=1}^{n} p(j) \sum_{i=1}^{n} p(i|j) a_j$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} p(i, j) \log \frac{p(i, j)}{p(j)} a_j$$

$$= -H(X, S) + H(x) + \sum_{j=1}^{n} p(j) \log a_j.$$

The difference between the two gains is

$$H(X) - H(X, S) + \Sigma p(j) \log a_j + H(s) - \Sigma p(j) \log a_j = H(X) + H(S) - H(X, S),$$

which by definition is the mutual information of X about S, denoted $I(X, S)$.

Hence the information processor which provides the most information also maximizes the a age rate of gain in resources. In general, we can define a measure of effectiveness of a proces $\underline{P}$ as a function of the prior $\underline{p}$ and the payoff odds $\underline{a}$ as

$$M(\underline{P}|\underline{p}, \underline{a}) = E_{\underline{p}}(G_{\max}(\underline{p}|\underline{X}, \underline{a})) - G_{\max}(\underline{p}, \underline{a}) = I(\underline{X}, \underline{S}|\underline{p}).$$

It is important to note, however, that the information provided by a processor, and hence measure of effectiveness, is a function of the prior distribution $\underline{p}$. A processor which is optimal one $\underline{p}$ might be inferior for another, as illustrated in the following example.

Given the prior $\underline{p}' = (0.5, 0.25, 0.25)$ and the transition matrices described above, we can culate the expected value of the log of the ratio of resources after processing and action to the p resources.

These calculations (see Appendix) yield

$$\log^{-1} M(I|\underline{p}', \underline{a}) = 1.63$$

and

$$\log^{-1} M(II|\underline{p}'', \underline{a}) = 1.78.$$

Hence processor II is to be preferred if $\underline{p} = (0.5, 0.25, 0.25)$. Note that processor II is more reli when $S = s_1$ is the correct state, and this happens half the time under $\underline{p}'$.

Suppose, however, that an attack from the side is more likely, say $\underline{p}'' = (0.2, 0.4, 0.4)$. Repea the calculations for this prior gives

$$\log^{-1} M(I|\underline{p}'', \underline{a}) = 1.62$$

$$\log^{-1} M(II|\underline{p}'', \underline{a}) = 1.52,$$

rocessor I is preferable.

n Section III, we give an example of an anti-missile missile allocation problem which gives a pos- interpretation of the "gain" referred to above. To avoid being distracted by the oversimplifica- of that example, we next discuss the second and third points listed above, concerning reaction to ges in the odds and comparisons of suboptimal strategies, in a simple gambling problem as set below.

EXAMPLE I. *A Three Letter Roulette Wheel.* Suppose that a gambler has available to him three ole wagers on three mutually exclusive and exhaustive events A, B, C, all at odds of 3 to 1; that s some device (sensor) which enables him to determine the relative probability of the three events; hat he has determined these probabilities to be 0.5, 0.3, and 0.2, respectively. If the gambler's ne is \$10, the *long-term* optimal strategy described here says he should bet \$5 on A, \$3 on B, 2 on C. In classical terms, his expectation on one play would then be

$$(0.5 \times 15 + (0.3) \times 9 + (0.2) \times 6 = \$11.40.$$

is less than the expectation he would have if he always bet his whole fortune on A, which would expected value $(0.5 \times 30 = \$15$. Nevertheless, in the long run, the latter greedy strategy leads st surely to ruin since it calls for always betting *all* of his fortune on the most likely outcome, and ually one of the less likely outcomes would occur and all would be lost. What appears to be good s would be poor *strategy!* A rational man might be tempted to hold back some of his resources never to be caught emptyhanded on the next round. The optimal strategy, however, seems to say t hold back anything, just bet in proportion to the probability."

This is a mathematically convenient way of describing the optimal allocation, but there is an alent strategy that involves partial withholding. An elementary calculation shows the gambler l get the same distribution by holding \$6 in reserve and betting \$3 on A and \$1 on B. Note, however, n either case he bets something on B, even though the payoff odds of 3 to 1 are less than $1/0.3 = 3.\underline{3}$. ssical gambler would never take such a bet because its expectation is negative. Nevertheless, in the withholding interpretation, intuition must yield to analysis, and the gambler should bet B. If he chooses to ignore the calculations and insists on only betting on favorable odds, the he can do is to withhold \$7.50 and bet \$2.50 on A, assuming such fractional bets are allowed. ı we calculate the effective rate of return of this strategy, we get

$$G(\underline{p}, \underline{a}, \underline{\alpha}) = 1/2 \log 15 + 1/2 \log 7.5 - \log 10,$$

sponding to a compound interest rate of 6.1 percent, whereas the return using the optimal strategy

$$G_{\max}(\underline{p}, \underline{a}) = 0.5 \log 15 + 0.3 \log 9 + 0.2 \log 6 - \log 10,$$

sponding to a compound interest rate of 7.14 percent.

Next, we can consider the consequences of a change in the probabilities to $\underline{p}' = (1/2, 1/3, 1/6)$.

The optimal strategy now calls for an allocation of \$5 to A, \$3.33 to B, and \$1.67 to C, which w give a return rate of 9.1 percent. But if house rules require whole dollar bets, the man with \$10 r choose to make the same \$5, \$3, and \$2 bets as before. Then, his expected return with this str would only be 8.6 percent. (In the example in Section III, such a house rule would correspo firing an integer number of missiles). Of course, another gambler in the same casino, with the information about the probabilities and \$6, could make the optimal proportional allocation of \$ and \$1. One difference between a casino and a tactical engagement is that the gambling casino w be as willing to cover a total bet of \$10 as it would \$6, but an aggressive adversary need not accommodating. This feature will come up again in the missile allocation problem.

Finally, we consider the effect on the allocation if the house were to change the payoff odds f letters from 3 to 1, to 5 to 3 on A and 5 to 1 on B and C. The gambler still knows the true probab are 0.5, 0.3, and 0.2 respectively. The odds satisfy the criterion of equitability ($\underline{p}' = (0.6, 0.2,$ so the optimal allocation is the original allocation of \$5, \$3, and \$2 despite the change in odds, sinc optimal allocation is proportional to the *true* probabilities which have not changed. What does ch is the effective rate of gain, which becomes 3.04 percent under the new odds. We will encounte property again in the missile allocation example.

III. A MISSILE ALLOCATION PROBLEM

In this section we consider a model of a tactical decision problem that has been simplifi satisfy the criterion of "equitability" and illustrate the use of the measure of effectiveness an evaluation of suboptimal strategies. To the extent that it is a gross oversimplification, the analy more qualitative than quantitative, but the principles are applicable to more complicated mode

We suppose that an adversary has fired a salvo of missiles at a task force, each missile capa inflicting damage whose value is four times the cost of the missile. The defender, however, ha kinds of anti-missile missiles both equal in cost to the attackers missiles. One, a quick respons face to Air Missile (SAM), can eliminate them on a one-to-one basis. The other is a multiple-inde ent-warhead Anti-Missile Missile (AMM) which can neutralize four of the incoming missiles but re more lead time. Specifically, if it is known in which sector the attack will arrive, then the m warhead missiles can all be fired off on general intercept trajectories and will eliminate the t Suppose, however, that the defender only knows that the attack is twice as likely to be coming in front than from either flank. Then the optimal strategy derived in Section II would allocate h the AMM's to the first sector and one quarter of them each to the other sectors.

Now the actual attack takes place in only one sector, so the missiles fired into the other secto fall wasted into the sea, but the missiles fired into the correct sector, whatever it is, will yield a return in relative cost. Thus this model is equivalent to the simplified roulette wheel exampl sidered above. Of course, the "winnings" in this case are relative, since it is the loser who i hind" rather than the winner who is "ahead."

Another less explicit assumption in this model is that the defender knows exactly how man siles have been fired in the salvo, so that he doesn't launch more AAM's than needed. This assun is generated by the fact that in gambling, the house will cover a bet exactly. The defender s launch just enough missiles to neutralize the attack and allot them to sectors according to the o strategy. To launch more missiles than necessary is equivalent to playing in a casino where th

e take. The 3 to 1 payoff on each of the three sectors is an "equitable" structure and corresponds ior distribution where the attack is equally likely to be in any one of the three sectors.

the prior intelligence corresponds to the vector $\underline{p}' = (1/2, 1/4, 1/4)$ as given above, the ex- gain from the optimal strategy is

$$G_{\max}(\underline{p}, \underline{a}) = \log 3 + (1/2) \log (1/2) + 2(1/4) \log (1/4)$$
$$= \log 1.061,$$

ponding to a 6 percent increase in "missile units." If processor *I* were available in this situation, d be expected to improve this gain to 63 percent.† Similarly, processor *II* would yield 78 percent. e other hand, with the prior $\underline{p}'' = (0.2, 0.4, 0.4)$, processor *I* would have the higher gain, as dis- in Section II.

o illustrate a comparison of strategies, suppose the prior distribution is $\underline{p} = (0.5, 0.3\underline{3}, 0.1\underline{6})$ salvo of 24 attacking missiles has been fired. The defender should fire three of his multiple war- nissiles into sector 1, two into sector 2, and one into sector 3 or, alternatively and equivalently, o sector 1 and one into sector 2, analogous to the gambling problem in the last part of Section II. er case his expected yield is 9 percent. Again, from Section II, if there were less prior informa- ay $\underline{p} = (0.5, 0.3, 0.2)$, and an integer missile allocation were required, this gain would drop to ent.

milarly, as discussed in the roulette example, if the antimissile effectiveness were different for and flank attacks, say three AAM's to counter five missiles in a frontal attack but only one to counter five missiles in a flank attack, then the allocation for $\underline{p} = (0.5, 0.3\underline{3}, 0.1\underline{6})$ would , and 1, as above, to counter a salvo of 15 missiles.

EMARKS ON ADAPTABILITY, RELIABILITY, AND INEQUITABILITY

possible conclusion from the comparison of processors *I* and *II* for priors $\underline{p}'$ and $\underline{p}''$ in the pre- section is that it would be convenient to have a switch that would enable us to select whichever sor characteristics were appropriate for the current prior. What this amounts to, of course, is an ıtary kind of adaptive processor which could be represented mathematically by a mapping from ace of vectors $\underline{p}$ to the space of $n \times n$ matrices. For any such adaptive processor and a specified of $\underline{p}$, it would still be possible within the scope of this theory to compute the expected gain of ocessor. Moreover, by putting a prior distribution on $\underline{p}$, for example a Dirichlet prior, one could te an overall expected gain for the adaptive processor.

a similar manner, the effect of reliability can be represented as a weighted average of processor es characterizing the various states of operation of the processor, including the percentage of ıe processor might be inoperative, and allocation of resources would be made on the basis of the

nally, it is clear that it might be too much to hope that there would be many situations that exactly criterion of equitability as given here. The theory developed by Kelly, Breiman et al., is by no restricted to what we have called equitable decision problems; it is just easiest to explain for se. But even for situations that we would call inequitable, corresponding to gambling in a casino there is a house take, it may still be the case that information, say the bias of a roulette wheel,

† See Appendix for calculations.

can be used to advantage. Here also the optimal strategy may call for taking some "bad" bets. the other hand, it is conceivable that the situation will be so inequitable that there will be no ad tageous bets or so superequitable that all bets are favorable. Even in these cases, it is important t aware that there are optimal allocation strategies derived from the extended theory.

The extended theory offers some suggestions for further study of information processors beca it indicates that there are fundamental differences between the optimal strategies to be emplo depending on whether a situation is "less than equitable" or "more than equitable." In the for situation, "bold" actions are advocated in the hopes of restoring equilibrium or gaining an advant since cautious policies would assure eventual total loss through slow attrition. But these same caut policies become the optimal strategies when a situation is "more than equitable," since such poli will, over time, preserve and increase the advantage.

It is very important, therefore, to know whether a situation is "more than equitable" or "less t equitable." But the ability to determine the type of situation depends largely on the extent and curacy of intelligence information about enemy weapon capabilities and the probability of attack various sorts, which must be obtained from sensor information processing. And since the opti allocation of resources depends on probabilities calculated from the operating characteristic ma of the processor, it is important to know what effect errors in the entries in these matrices would h on the outcome. It would seem that further study of this aspect of the problem is worthwhile.

V. CONCLUSIONS AND RECOMMENDATIONS

We wish to stress that the measure of effectiveness proposed here is not offered as a yardstick all occasions. It was derived for a certain kind of recurring situation, and its application to sensor strategies is only suggested for these situations. There are certainly other types of situations for wh an information-based measure would be inappropriate. Indeed, Schwartz [4] concludes that for a ra in a Distant Early Warning line type of installation, reliability, not information rate, is the pro measure.

The principal points of this paper, then, are the following:

(i) There can be situations in which a measure of effectiveness of a data processor based on formation rate has, beyond its intuitive appeal, a quantitative interpretation in terms of growth rat relative resources.

(ii) This measure is calculated with reference to an optimal strategy for using the informat and thus also provides a means for comparing other suboptimal strategies.

(iii) A surprising, nonintuitive result of the analysis is that in some situations taking "bad" b may be good strategy.

(iv) These conclusions carry over in a qualitative way to situations that are not equitable, but nature of optimal strategies depends distinctly on the type of departure from equitability.

VI. ACKNOWLEDGMENTS

The author is grateful to Ray Hershman of Code 3400 at the U.S. Navy Electronics Laborat Center (NELC), San Diego, for suggesting this study as a result of discussions about some shortcomi of a measure of effectiveness proposed by Planning Research Corporation [5].

Mr. Hershman and his colleagues at NELC were also helpful in suggesting the kind of tact situation in which the proposed MOE might be appropriate.

APPENDIX

n this appendix, we calculate the expected value of the log of the gain for processors *I* and *II* e prior distributions $\underline{p}'$ and $\underline{p}''$ given in Section II.

Recall the matrix for the operating characteristics for processor *I* was

		OUTPUT X		
		1	2	3
TRUE STATE	1	0.8	0.1	0.1
	2	0.1	0.8	0.1
	3	0.1	0.1	0.8

xample, if $S = s_1$ is the true state, then $P(X = 1 \mid S = s_1) = 0.8$, etc. Multiplying the prior vector (0.5, 0.25, 0.25) times this matrix gives the distribution of the output X as

$$P(X = 1) = 0.45,\ P(X = 2) = P(X = 3) = 0.0275.$$

an elementary application of Bayes' formula gives

$$(\underline{p} \mid X = 1) = (0.8\underline{8},\ 0.0\underline{5},\ 0.0\underline{5}),$$

$$(\underline{p} \mid X = 2) = (0.\underline{18},\ 0.\underline{72},\ 0.\underline{09}),$$

$$(\underline{p} \mid X = 3) = (0.\underline{18},\ 0.\underline{09},\ 0.\underline{72}).$$

he optimal allocation when $X = 1$ is therefore $\underline{\alpha} = (0.8\underline{8},\ 0.0\underline{5},\ 0.0\underline{5})$, and

$$\begin{aligned} G_{\max}(\underline{p} \mid X = 1, \underline{a}) &= \log 3 + 0.8\underline{8} \log (0.88) + 2(0.05\underline{5}) \log (.05\underline{5}) \\ &= \log 1.9596. \end{aligned}$$

the expected gain is about 96 percent when $X = 1$, and this happens about 45 percent of the time. imilarly for $X = 2$ or $X = 3$, we have

$$G_{\max}(\underline{p} \mid X = 2, \underline{a}) = G_{\max}(\underline{p} \mid X = 3, \underline{a}) = \log 1.4036,$$

e are in one or the other of these situations a total of 55 percent of the time. Hence

$$\begin{aligned} E(G_{\max}(\underline{p} \mid X, \underline{a})) &= 0.45 \log 1.9596 + 0.55 \log 1.4036 \\ &= \log 1.631. \end{aligned}$$

, with prior $\underline{p} = (0.5,\ 0.25,\ 0.25)$, and before we see the value of X, the average expected gain proximately 63 percent. To relate this to information theory, we note that the prior entropy is

1.5 bits and the expected posterior entropy is $\log_2 1.631 = 0.879$ bits, so the processor can be expec to provide 0.621 bits in this situation.

Repeating the calculation for the second processor, whose matrix was

		OUTPUT X		
		1	2	3
TRUE STATE	1	0.9	0.05	0.05
	2	0.1	0.7	0.2
	3	0.1	0.2	0.7

we find that $E(\underline{X}) = (0.5, 0.25, 0.25)$ and

$$(\underline{p}\,|X = 1) = (0.9, 0.05, 0.05),$$

$$(\underline{p}\,|X = 2) = (0.1, 0.7, 0.2),$$

$$(\underline{p}\,|X = 3) = (0.1, 0.2, 0.7).$$

Also

$$G_{\max}\,(\underline{p}\,|X = 1, \underline{a}) = \log 2.35,$$

$$G_{\max}\,(\underline{p}\,|X = 2, \underline{a}) = G_{\max}\,(\underline{p}|X = 3, \underline{a}) = \log 1.345,$$

$$E\,(G_{\max}\,(\underline{p}\,|X, \underline{a})) = \log 1.778.$$

Hence the expected gain using the second processor is approximately 78 percent, and it wo be preferred for the prior $\underline{p}'$. When we evaluate the information we can expect the second proces to provide in this case, we find the expected gain in information is 0.745 bits, which is more t expected from the first processor.

However, when we compare the two processors for the prior $\underline{p}'' = (0.2, 0.4, 0.4)$, we find that processor *I*

$$(\underline{p}\,|X = 1) = (0.66, 0.16, 0.16),$$

$$(\underline{p}\,|X = 2) = (0.053, 0.842, 0.105),$$

$$(\underline{p}\,|X = 3) = (0.053, 0.105, 0.842),$$

and $E\,(G_{\max}\,(\underline{p}\,|X, \underline{a})) = \log 1.62$. For processor *II* we have

$$(\underline{p}\,|X = 1) = (0.692, 0.154, 0.154),$$

$$(\underline{p}\,|X = 2) = (0.027, 0.757, 0.216),$$

$$(\underline{p}\,|X = 3) = (0.027, 0.216, 0.757),$$

and $E\,(G_{\max}\,(\underline{p}\,|X, \underline{a})) = \log 1.52$.

or this second prior, the average gain of 62 percent for processor *1* is less than it was for the ior, but more than the gain from the second processor, which is only 52 percent in this case.

REFERENCES

eiman, L., "Optimal Gambling Systems for Favorable Games," Proceedings of the Fourth Berkeley Symposium in Probability and Mathematical Statistics *1*, 65–71 (1961).

bins, L. E. and Savage, L. J., "Optimal Gambling Systems," Proceedings of the National Academy of Science, U.S.: 1597 (Oct. 19, 1960).

lly, J. L., "A New Interpretation of Information Rate," Bell Syst. Tech. J. *35*, 917–926 (July 956).

hwartz, L. S., "Marginal Utility and a Criterion of Performance for Communication and Radar Systems," Proceedings of Institute of Electrical Engineers (May 1959), p. 117–119.

ystem Design Considerations for a Bayesian Anti-Sub Submarine Warfare Information Processor," Planning Research Corporation Report R–1319 (April 1969).

AN INFILTRATION GAME WITH TIME DEPENDENT PAYOFF

Marlin U. Thomas and Yair Nisgav*

Naval Postgraduate School
Monterey, California

ABSTRACT

The problem of assigning patrol boats, subject to resource constraints, to capture or delay an infiltrator with perishable contraband attempting escape across a long, narrow strait is formulated as a two-sided time sequential game. Optimal mixed strategies are derived for the situation of one patrol boat against one smuggler. Procedures for obtaining numerical solutions for $R > 1$ patrol boats are discussed.

RODUCTION

is paper describes an application of game theory for examining strategies available to a patrol rsuing smugglers of perishable items who attempt escape by crossing a long, narrow strait. ber of applications of game theory to military-type problems have been reported. Recently, er and Payne [2] discussed an application of two-sided games in examining logistics allocation ns in a combat setting. Charnes and Schroeder [1] have developed some models of tactical ns in Antisubmarine Warfare. More recently, Pugh [5] has discussed some time-sequential rson zero sum games for treating strategic and tactical decisions within a given time frame. application described here, we formulate a twosided time-sequential game where one side, rol unit, has limited resources to catch his opponent, an infiltrator, who must make his escape a fixed time period.

ATROL GAME.

e consider a long, narrow strait where smuggling activity is taking place. Let side A represent unit whose objective is to capture or reduce the value of contraband held by side B, the infiltra- muggler seeking escape by crossing the strait to exit from side A's territory. The contraband side B is perishable with a lifetime of M time units; consequently, he must make his escape M time units in order to benefit from his infiltration. An example of the type of contraband is ence information. Side A is under a single command equipped with speedboats containing radar and communication units. Side B is an individual unit with small motorboats. Although has search radar, due to the narrowness of the strait, side B's radar echo will be shadowed , thus making radar detection near the shore virtually impossible. Thus, A can detect B only sufficiently far from shore. For obvious reasons, side B only attempts escape at night, and he

utenant, Israeli Navy.

departs from a point near a village or parallel to a village located on the other side of the stra though the patrol boats are much faster than B's, the fact that the strait is long and narrow give B a chance to cross successfully without being detected.

Viewed as a game, both sides would like to use their "best" strategies. The best strategy fo that which maximizes the number of boats captured from side B. Side B views his best strategy a one maximizing the number of trips per boat before capture. We make the following further si fying assumptions.

Assumptions:

1. Detection information for side A is perfect in the sense that there are no errors, and once detected, B is caught.
2. A's resources are limited to $k < M$ night patrols.
3. B's success requires a single crossing of the strait during the period of M nights.
4. Both A and B know the values of k and M.

The first assumption is merely for simplification to limit the scope of the problem. As we poir later, relaxing this assumption requires only a slight extension. Implicitly, we are assuming that light traffic exists in the channel, as one would expect for a channel that is being patrolled. Sinc boats are much faster, once B is a sufficient distance from shore to be detected, he can neither nor return fast enough if, in fact, he is detected. We are also excluding from our consideration t boats belonging to side B that are crossing into A's territory. Although the second assumption seem somewhat superficial, such problems of limited resources are becoming realities for many mil components.

2.1. One Patrol Boat–One Smuggler

First, we shall consider the situation where each side has a single boat. Side B decides for e night to go or not to go and attempt escape across the channel, while A similarly makes a decisio whether or not to have a patrol in the channel.

We shall denote by $\Gamma(n, k)$, $k < M$, $n = M, M-1, \ldots, 1$, the game determined by the ab assumptions for the nth day before the end of the period M. The game matrix for this game is as foll

A	B Smuggler's Actions	
	Go	No go
Assign patrol	v Game over	$\Gamma(n-1, k-1)$
No patrol	-1 Game over	$\Gamma(n-1, k)$

Whenever side A assigns a patrol and B decides to attempt escape, side A has some probabilit catching side B. This probability can be determined by solving a zero sum game [3]. If side B dec to go and A does not have a patrol out, then B wins the game and we assign to A, for convenie

off of -1. Likewise, we let the ultimate payoff to A for capturing B be $+1$. The remaining alves result in a loss of one available day, which is of benefit to side A. If side A assigns l and B does not go, then both sides face the game $\Gamma(n-1, k-1)$. If A chooses not to assign a while B elects not to go, then they face the game $\Gamma(n-1, k)$.

et $g(n, k)$ represent the value of the game $\Gamma(n, k)$. It follows (from [6, p. 173]) that $g(n, k)$ is by the recursive relationship

$$g(n, k)=\frac{v\cdot g(n-1, k)+g(n-1, k-1)}{v+g(n-1, k+1)-g(n-1, k-1)}$$

ie boundary conditions $g(n, 0)=-1$ and $g(n, n)=v, \forall n>0$. We note that for the last period , the game matrix is

$$\Gamma(1, 0)=\begin{bmatrix} v & 1 \\ -1 & 1 \end{bmatrix}$$

$|v|<1$. By dominance, side B must choose the go strategy, which implies that he always elects n the last period if he has not attempted escape before. Now if side A has $k=n$ available nights to patrols, then clearly he will use them all. Hence, the value of the game is v, since we know that go on one of these nights. If side A has $k>n$ available nights for patrol, then with a single patrol e can assign at most n of them.

HEOREM 2.1: The solution to the difference equation (1) for the game $\Gamma(n, k)$, $k<M$, $n=M$, . . ., 1 is

$$g^*(n, k)=\frac{k(v+1)-n}{n}.$$

ROOF: The proof follows directly by substituting (1) into (2) and simplifying.

e can now apply this result to our game matrix to obtain

$$\Gamma(n, k)=\begin{bmatrix} v & \dfrac{(k-1)(v+1)-(n-1)}{n-1} \\ -1 & \dfrac{k(v+1)-(n-1)}{n-1} \end{bmatrix}$$

otimal mixed strategies for A and B can be determined from (3). Let x_k^n be the probability that "assign a patrol" and y_k^n the probability that B will "go" when n nights remain and A's resources patrols until the end of the period. It follows that the optimal choices for these probabilities en by

$$x_k^n=k/n, \ (k<M;\ n=M, M-1, \ldots, 1)$$

$$y_k^n=1/n, \ (k<M;\ n=M, M-1, \ldots, 1).$$

We conclude that in order to obtain the value of the game, side A must allocate his available that he can assign patrols such that his probability of assigning a patrol is equal to the ratio number of search periods available to him to the total number of remaining periods. For side B form distribution over the remainder of the time period will provide him the value of the game. in particular, that this probability, y_k^n, does not depend on the number of available days that A h assigning patrols.

EXAMPLE: In order to demonstrate these results, consider the game $\Gamma(n, k)$, whereby period if side A allocates a patrol when side B has elected to "go," then A receives a payoff Suppose further that 10 nights remain until the end of the period, but A has only six nights av to him for assigning patrols.

Thus, we have $v=0.5$, $n=10$, and $k=6$ for which we get from (4a) and (4b) that $x_6^{10}=0.6$ and y_6^{10} and from (2) that the value of the game is $g^*(10, 6)=-0.1$. Now if A did not "assign a patrol" did not "go," then A and B face the new game $\Gamma(9, 6)$ for which: $x^9{}_6=0.667$, $y^9{}_6=0.112$, and $g^*(9, 6$

2.2. Two Patrol Boats—One Smuggler

Unfortunately, we do not have such closed form results for situations where A and B have than one boat. We shall, however, discuss formulations for deriving numerical solutions when two patrol boats. Let $k_1 < M$ and $k_2 < M$ represent the number of patrols that can be assigned to boats due to limited resources. There are two cases to be considered.

Case 1—Identical Patrol Boats

Suppose the two patrol boats are identical and A can make assignments of patrols accord some optimal plan. Let V_1^* and V_2^* be payoffs to A for assigning one and two patrols, respectively given night when B chooses to "go". Side A now has three alternatives, and the game matrix the form

$$\Gamma(n, k)=\begin{bmatrix} V_2^* & \Gamma(n-1, k-2) \\ V_1^* & \Gamma(n-1, k-1) \\ -1 & \Gamma(n-1, k) \end{bmatrix}, \tag{5}$$

with the boundary conditions leading to

$$\Gamma(n, 1)=\begin{bmatrix} V_1^* & -1 \\ -1 & \Gamma(n-1, 1) \end{bmatrix} \quad \text{and} \quad \Gamma(1, k)=\begin{bmatrix} V_2^* & 1 \\ V_1^* & 1 \\ -1 & 1 \end{bmatrix},$$

noting that on the last day of the period

$$g(1, k)=\begin{cases} V_2^* & \text{if } k \geq 2 \\ V_1^* & k=1 \\ -1 & k=0. \end{cases}$$

t that the two patrol boats are identical allows us to lump together the remaining available for the period. The solution to this game can be derived through a recursive equation of the

$$g(n, k)=f(V_1^*, V_2^*, \Gamma(n-1, k), \Gamma(n-1, k-1), \Gamma(n-1, k-2)),$$

. This calls for the solution of a 3×2 game, which typically is solved by linear programing. We it for this particular structure, the dual to the standard LP problem is

$$\begin{aligned} &\min W \\ \text{s.t. } &V_2^*\cdot y+g(n-1, k-2)\cdot(1-y)\leq W \\ &V_1^* y+g(n-1, k-1)\cdot(1-y)\leq W \\ &-y+g(n-1, k)\cdot(1-y)\leq W. \end{aligned}$$

r to have dominance, it is necessary that $V_2^*\geq V_1^*\geq -1$ and $g(n-1, k)\geq g(n-1, k-1)$ $-1, k-2)$. The LP given by (6) can conveniently be solved graphically.

–Nonidentical Patrol Boats

nsider now the case where the two patrol boats are not identical. Let V_1^*, V_2^*, and V_3^* represent ected payoffs when boat number 1, boat number 2, and both boats, respectively, are assigned according to some optimal allocation procedure. Denote by $\Gamma(n, k_1, k_2)$ our game played when ls remain and side A has k_1 available patrols for patrol boat number 1 and k_2 for boat number 2. ne matrix then is

$$\Gamma(n, k_1, k_2)=\begin{bmatrix} V_3^* & \Gamma(n-1, k_1-1, k_2-1) \\ V_2^* & \Gamma(n-1, k_1-1, k_2) \\ V_1^* & \Gamma(n-1, k_1, k_2-1) \\ -1 & \Gamma(n-1, k_1, k_2) \end{bmatrix},$$

boundary conditions

$$\Gamma(1, k_1, k_2)=\begin{cases} V_3^*, & \text{if } k_1\geq 1, k_2\geq 1 \\ V_2^*, & k_1\geq 1, k_2=0 \\ V_1^*, & k_1=0, k_2\geq 1 \\ -1, & k_1=0, k_2=0 \end{cases}$$

$$\Gamma(n, 0, 0)=-1.$$

ıtion to this game can be obtained numerically using the procedures described for Case 1.

NCLUDING REMARKS

with any model, the models presented here are mere abstractions of reality. Thus, the major rovided are through insights from identifying and examining various relationships among

operational parameters. It is of interest in maintaining a patrol capability (side A) to know w happen if certain conditions are changed. In particular, one is concerned with how side A's effect and best strategy vary if he changes the number of patrol boats available for assignment to th

There are a host of extensions that could be made to the present study. In principle, the s where side A has $R > 2$ patrol boats when B has one boat can be treated in a similar fashion as i.e., with two nonidentical patrol boats, only the game matrix is $2^R \times 2$. One must define all combi of payoffs and numerically solve a recursive relationship for the value of the game.

A much more difficult problem, but indeed one of interest, is the case where side B has mo one boat. Depending upon the payoffs involved, it might be more reasonable from B's point of send out a number of boats, some of which are missioned to deceive or confuse A. From A's vie this problem begins to take the form of a type of search and detection problem (see Pollock [4]) present study we assumed that side A had a constant detection capability and that perfect infor was gained through detection. Although admittedly it is a more difficult problem computat conceptually one can extend these games to allow for false alarms.

REFERENCES

[1] Charnes, A. and R. G. Schroeder, "On Some Stochastic Antisubmarine Games," Nav. Re Quart. *14*, 291–311 (1967).

[2] Moglewer, S. and C. Payne, "A Game Theory Approach to Logistics Allocation," Nav. Re Quart. *17*, 87–97 (1970).

[3] Nisgav, Y., "Some Game Theory Models for Allocating Forces in a Narrow Strait Against Activity," M.S.O.R. Thesis, Naval Postgraduate School, Monterey, California (Sept. 1973

[4] Pollock, S. M., "Search Detection and Subsequent Action: Some Problems on the Inter Operations Research *19*, 559–586 (1971).

[5] Pugh, G. E. and J. P. Mayberry, "Theory of Measures of Effectiveness for General-Purpose M Forces: Parts I, II," Operations Research *21*, 867–906 (1973).

[6] von Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Princeton, New Jersey, 1947).

THE BILINEAR PROGRAMMING PROBLEM

Harish Vaish
California State University, Northridge
and
C. M. Shetty
Georgia Institute of Technology

ABSTRACT

The paper deals with bilinear programming problems and develops a finite algorithm using the "piecewise strategy" for large-scale systems. It consists of systematically generating a sequence of expanding polytopes with the global optimum within each polytope being known. The procedure then stops when the final polytope contains the feasible region.

RODUCTION

e Bilinear Programming Problem discussed in this study may be stated as:

$$\begin{aligned} \text{BLP: Minimize} \quad & \phi(x, y) = c^t x + d^t y + x^t C y \\ \text{subject to} \quad & x \in X_0 = \{x \in R^m \mid Ex \leq e,\ x \geq 0\} \\ & y \in Y_0 = \{y \in R^n \mid Gy \leq g,\ y \geq 0\}. \end{aligned}$$

t loss of generality, we will assume that X_0 and Y_0 are bounded polytopes. In spite of its special re, problem BLP is a mathematical statement of a number of practical problems (see [5] and It is also closely related to a general quadratic programming problem where the objective funcnot necessarily convex.

this point it is worth noting that a BLP problem has an interesting and useful property, namely mum is attained at an extreme point, as asserted by Theorem 1 below.

HEOREM 1: Suppose X_0 and Y_0 are nonempty and compact. Then the global minimum of BLP ned at $(\bar{x}, \bar{y})$, where $\bar{x}$ is an extreme point of X_0 and $\bar{y}$ is an extreme point of Y_0.

ROOF: Trivial by noting that $\phi(x, y)$ is linear in x for fixed y and linear in y for fixed x.

wever, it can be shown that ϕ is not explicitly quasi-convex, so that local minima can and do 0]. It is this aspect of the problem that causes the essential difficulty in the solution procedure. veral methods for solving such problems have been proposed in the literature, including cutting lgorithms. Konno [5] has developed a cutting plane algorithm for problem BLP which consists tioning the set X_0 (or Y_0) into subsets such that the global minimum is known over each subset $y \in Y_0$ (*or* $x \in X_0$). The method converges to an ϵ-optimal solution. This study deals with a polyannexation strategy where we have the global minimum over a sequence of polytopes which are d until the feasible region is contained in a final polytope. Tui [9] was the first to propose such an

algorithm to solve the problem: Minimize $f(x)$ subject to $x \epsilon D \subset R^n$ where D is a polyhedron and concave. Let $\bar{x}$ be a nondegenerate local star minimum, and let $y^1, y^2, \ldots, y^n$ be points on the n r incident on $\bar{x}$ and furthest from it such that $f(y^i) \geq f(\bar{x})$. Let $P_0 = \text{Conv}\,[\bar{x}, y^1, y^2, \ldots, y^n]$. If $D \subset$ the problem is solved. If not, an extreme point x^1 of D furthest from the hyperplane H passing thro $y^1, y^2, \ldots, y^n$ is found. This point is "projected" to a point y^{n+1} along the ray joining $\bar{x}$ to x^1 such t $f(y^{n+1}) \geq \min\{f(\bar{x}), f(x^1)\}$. Let $y^{n+1} = \sum_{i=1}^{n} \alpha_i y^i$. For each $\alpha_i \neq 0$, a new hyperplane H_i is defined replacing y^i with y^{n+1}, the other points being the same. The polytope P_0 is expanded to P_1, which cludes y^{n+1}. The procedure terminates when $D \subset P_r$ for some polytope P_r in the sequence of polyto generated.

Without providing a formal proof, Tui asserted that the method could be shown to be finite. Zw [12] provided a counterexample to show that this is not true, and proposed a modified polytope anne tion algorithm which converges to only an ϵ-optimal solution. The modification insures that the po x of D found at any stage satisfies $\beta_i \geq 0$ where β_i is defined by $x^r = \sum_{i=1}^{n} \beta_i y^k i$ and $y^k i$ are the poi defining the hyperplane under consideration. In this case x^r need not be an extreme point of D. Furth to insure finite convergence, Zwart generates additional hyperplanes only if the distance of x^r from t hyperplane is greater than some specified $\epsilon > 0$ leading to an ϵ-optimal solution.

Gallo and Ülkücü [3] modified Tui's algorithm to solve problem BLP. The essential difference fr Tui's method is that a new hyperplane is generated corresponding to each $\alpha_i > 0$ instead of $\alpha_i \neq 0$. T modification is sufficient to ensure that the point y^i just replaced will not be located in the immedi next stage. However, as shown in [10] by an example, the modification is not adequate to ensure t y^i will never be located again. Hence this method also leads to cycling.

Incidentally, Shachtman [7] considers the problem of optimizing a linear function over a set points $S = \{v^1, \ldots, v^n\}$, $v^i \epsilon R^n$. Shachtman uses Tui's method to generate the polytope $P = \text{co}$ $[v^1, \ldots, v^n]$ and then optimizes the function over P. Shachtman asserts that the procedure is fini However, as mentioned in a private communication [8], special programing aids were needed to ensu convergence.

The purpose of this paper is to present a finitely convergent polytope annexation algorithm f solving BLP. With minor modifications, it can also be applied to a concave minimization proble Section II presents the methodology for inductively constructing all the facets of a finite sequence bounded polytopes. This is used in Section III to develop a polytope annexation algorithm for BL Some computational considerations are also discussed.

II. FACET GENERATION

The procedure discussed in this paper requires the generation of all the facets of a bounded po tope $P_{i+1} = \text{conv}\,[P_i \cup \{v\}]$, where $v \epsilon R^m$, $v \notin P_i$. This is discussed in detail later in this section a relies on the concept of whether a given point v is beneath or beyond a given hyperplane as discuss below.

Definition 1. Let P be a convex subset of R^m. A set F, $F \subset P$, is a *face* of P if either $F = \phi$, $F = P$, or if there exists a supporting hyperplane H of P such that $F = P \cap H$. ϕ and P are called t *improper* faces of P. All other faces are called *proper* faces.

Definition 2. A maximal proper face of P, that is, a proper face of the highest possible dimensio is called a *facet* of P. Thus if P is an m-dimensional polytope, a facet of P is $(m-1)$-dimensional.

efinition 3. Given a set A, the *affine hull* of A is the set aff $[A]=\{x|x=\sum_{i=1}^{i=k}\lambda_i x^i, \sum_{i=1}^{i=k}\lambda_i=1, x^i\epsilon A,$ . . ., $k\}$ for arbitrary finite k.

efinition 4. Given a set A, the *convex hull* of A is the set conv $[A]=\{x|x=\sum_{i=1}^{i=k}\lambda_i x^i, \lambda_i\geq 0, \sum_{i=1}^{i=k}$ $x^i\epsilon A, i=1, \ldots, k\}$ for arbitrary finite k.

efinition 5. Let P be an m-dimensional polytope and H be a hyperplane such that $H\cap$ int $(P)=\phi$. R^m where v is said to be *beneath* H (with respect to P) provided v belongs to the open halfspace ined by H which contains int (P). If F is a facet of P, v is beneath F if v is beneath aff $[F]$. ise, if v is in the open halfspace determined by H which does not contain int (P), then v is said eyond aff $[F]$.

EMMA 1: *Let* $P\subset R^m$ be a polytope, $F=$ conv $[x^1, \ldots, x^m]$ a facet of P, aff $[F]=\{x\epsilon R^m|p^tx=$ d $P\subset$ aff $[F]^-=\{x\epsilon R^m|p^tx\leq 1\}$. Let $v\epsilon R^m$, $v\sum_{i=1}^{i=m}\lambda_i x^i$. Then v is beyond aff $[F]$ if and only if >1, and v is beneath aff $[F]$ if and only if $\sum_{i=1}^{i=m}\lambda_i<1$.

ROOF: Let v be beyond aff $[F]$. Then $v\epsilon$ int (aff $[F]^+$). Hence $1<p^tv=\sum_{i=1}^{i=m}\lambda_i p^tx^i=\sum_{i=1}^{i=m}\lambda_i$.

rsely, let $\sum_{i=1}^{i=m}\lambda_i>1$. Then $p^tv=\sum_{i=1}^{i=m}p^t\lambda_i x^i=\sum_{i=1}^{i=m}\lambda_i p^tx^i=\sum_{i=1}^{i=m}\lambda_i>1$. Hence $v\epsilon$ int (aff $[F]^+$). $P\subset$ aff $[F]^-$, v is beyond aff $[F]$. The proof of the other part is similar.

e will now relate the above concepts to the question of determining the faces of the polytope P' d from a polytope P. Theorem 4 below gives the relationship between the facets of P' and P and principle theorem justifying the solution procedure.

HEOREM 2: If P is an m-dimensional polytope, each $(m-2)$-dimensional face F of P is contained isely two facets F_1 and F_2 of P, and $F=F_1\cap F_2$.

ROOF: See [4].

HEOREM 3: Let P and P' be two m-dimensional polytopes in R^m, and let v be a vertex of P', uch that $P'=$ conv $[\{v\}\cup P]$. Then

A face F of P is a face of P' if and only if there exists a facet F' of P such that $F\subset F'$ and v is h F'.

) If F is a face of P, then $F'=$ conv $[\{v\}\cup F]$ is a face of P' if and only if either (a) $v\epsilon$ aff $[F]$ or ong the facets of P containing F there is at least one such that v is beneath it and at least one hat v is beyond it.

oreover, each face of P' is of one and only one of those types.

ROOF: See [4].

HEOREM 4: Let P and P' be two m-dimensional polytopes in R^m, let v be a vertex of P', $v\notin P$, hat $P'=$ conv $[\{v\}\cup P]$. Then

A facet F of P is a facet of P' if and only if v is beneath aff $[F]$.

) Let F be a facet of P. Then $F'=$ conv $[\{v\}\cup F]$ is a facet of P' if and only if $v\epsilon$ aff $[F]$.

i) Let F be a $(m-2)$-dimensional face of P. Let F_1 and F_2 be the two facets of P containing F. $F'=$ conv $[\{v\}\cup F]$ is a facet of P' if and only if v is beneath aff $[F_1]$ (or aff $[F_2]$) and beyond] (or aff $[F_1]$).

(iv) Each facet of P' is either a facet of P or is of the form $F' = \text{conv}\,[\{v\} \cup F]$ where F is of P.

PROOF:

(i) Let F be a facet of P, and let v be beneath F. Then from part (i) of Theorem 3, F is a face But F is $(m-1)$-dimensional since it is a facet of P. Hence it is a facet of P'.

Conversely, let F be a facet of both P and P'. Then from part (i) of Theorem 3 there exists $\bar{F}$ of P such that F is contained in $\bar{F}$ and v is beneath $\bar{F}$. But F and $\bar{F}$ being facets of P' and implies $F = \bar{F}$. Hence v is beneath F.

(ii) Let F be a facet of P and $v \in \text{aff}\,[F]$. From part (ii) (a) of Theorem 3, $F' = \text{conv}\,[\{v\} \cup$ a face of P'. But F is $(m-1)$-dimensional, and hence F' is $(m-1)$-dimensional. Hence F' is of P'.

Conversely, let F be a facet of P, and let $F' = \text{conv}\,[\{v\} \cup F]$ be a facet of P'. Hence both F' are $(m-1)$-dimensional. Hence $v \in \text{aff}\,[F]$.

(iii) Let F be a $(m-2)$-dimensional face of P, and let v be beneath F_1 and beyond F_2. From (ii) (b) of Theorem 3, $F' = \text{conv}\,[\{v\} \cup F]$ is a face of P'. Now $v \notin \text{aff}\,[F_1]$ and $v \notin \text{aff}\,[F_2]$ s $v \notin \text{aff}\,[F_1] \cap \text{aff}\,[F_2]$. From Theorem 2, $F = F_1 \cap F_2$, and hence $\text{aff}\,[F] = \text{aff}\,[F_1] \cap$ aff Hence $v \notin \text{aff}\,[F]$. Hence F' is $(m-1)$-dimensional.

Conversely, let F be a $(m-2)$-dimensional face of P and F' be a facet of P'. Since F' is $(m$ dimensional, $v \notin \text{aff}\,[F]$. Also, from Theorem 2 there are precisely two facets F_1 and F_2 of P suc $F_1 \cap F_2 = F$. Hence from part (ii) (b) of Theorem 3, v is beneath F_1 (or F_2) and beyond F_2 (or F_1

(iv) From Theorem 3, each facet of P' is either a face of P or is of the form conv $[\{v\} \cup$ where F is a face of P. In the former case, F is a facet of P. In the latter case, F is $(m-1)$-dimen and $v \in \text{aff}\,[F]$, or F is $(m-2)$-dimensional and $v \notin \text{aff}\,[F]$.

Using Theorem 4, given a vertex $v \notin P_i$ and knowing all the facets of P_i, all the facets of P_{i+} be constructed. Let the facets of P_i be $\theta_{i1}, \ldots, \theta_{in}$. Let $x^k \notin P_i$. In order to construct all the fac $P_{i+1} = \text{conv}\,[P_i \cup \{x^k\}]$, the facets of P_i are sorted into three classes as follows:

(a) This will consist of all facets θ_{ie} of P_i such that x^k is beneath aff $[\theta_{ie}]$. From Theorem 4 (i), θ_{ie} will be a facet of P_{i+1}. If $\theta_{ie} = \text{conv}\,[x^{1,ie}, \ldots, x^{n,ie}]$ and $x^k = \sum_{j=1}^{j=n} \lambda_j x^{j,ie}$, then from Lem θ_{ie} will be in this class if $\sum_{j=1}^{j=n} \lambda_j < 1$.

(b) This will consist of all facets θ_{im} of P_i such that $x^k \in \text{aff}\,[\theta_{im}]$. From Theorem 4 part (ii), $[\theta_{im} \cup \{x^k\}]$ will be a facet of P_{i+1}. θ_{im} will be in this class if $\sum_{j=1}^{j=n} \lambda_j = 1$.

(c) This will consist of all those facets θ_{iq} of P_i such that x^k is beyond aff $[\theta_{iq}]$. From Lem θ_{iq} will be in this class if $\sum_{j=1}^{j=n} \lambda_j > 1$. In order to use Theorem 4, part iii, all $(m-2)$-dimensional F of P_i have to be identified such that there exists a facet θ_{ie} of P_i in class (a) which contains F facet θ_{iq} of P_i in class (c) which contains F. From Theorem 2, it is clear that there exist precisel facets of P_i containing F. This can be done by determining whether or not θ_{ie} and θ_{iq} have $(m$ elements in common. If they do, let these elements be $x^1, \ldots, x^{m-1}$. Then conv $[x^1, \ldots, x^{m-}$ a $(m-2)$-dimensional face of P_i which satisfies the hypotheses of Theorem 4, part (iii). Hence $[x^1, \ldots, x^{m-1}, x^k]$ is a facet of P_{i+1}. From Theorem 4, part (iv), the facets generated in this ma will be all the facets of P_{i+1}.

or the Bilinear Problem, finding a vertex x^k of X_0 which is not contained in the current polytope easy. Let $\theta_j = \text{conv}\,[\bar{x}^{j_1}, \bar{x}^{j_2}, \ldots, \bar{x}^{j_m}]$ be any facet of P_i. Let B_j be a $m \times m$ matrix with $\bar{x}^{j_i}$, $\ldots, m$ as its columns. Consider the following linear program (which is defined as LP2 later):

$$\text{Max } Z = eB_j^{-1}x \qquad \text{subject to } x \in X_0,$$

e is a row vector with each element equal to 1. Let $\bar{Z}$ be the maximum value of the objective on attained at x^q. If $\bar{Z} > 1$, then x^q is beyond aff $[\theta_i]$ and hence $x^q \notin P_i$. If $\bar{Z} = 1$, then $x^q \in$ aff $[\theta_i]$. case, it is not clear whether or not $x^q \in P_i$. One way of answering this question is to generate all ative optimal solutions to problem LP2 and compare them with the elements of θ_i. However, P_i, x^q will clearly be beyond some facet of P_i. Hence a simpler approach is to neglect all facets or which $\bar{Z} = 1$. If $X_0 \not\subset P_i$, at least one facet of P_i will give $\bar{Z} > 1$ and yield an extreme point of contained in P_i. Conversely, if $\bar{Z} = 1$ for each facet of some polytope P_r in the sequence of poly- then $X_0 \subset P_r$. This property is used in the algorithm of Section III.

OLYTOPE-ANNEXATION ALGORITHM

onno [5] has shown that any vertex of X_0 and Y_0 can be made the origin of the coordinate system ll maintain the structure of problem BLP. For ease of presentation, we will assume that the origin nondegenerate vertex of X_0, since the usual perturbation methods can be used to handle de- cy. We will also define the following problems which need to be solved at different stages of the hm:

I. LPI: $\underset{\lambda > 0}{\text{Max}} \{\underset{y \in Y_0}{\text{Min}}\, \lambda_c{}^t\hat{c} + d^t y + \lambda \hat{x}^t C y \geq k\}$

with $\hat{x}$ and k fixed.

I. LP2: $\text{Max } Z = eB_j^{-1}x$

subject to $x \in X_0$

with $B_j = [\bar{x}^{j_1}, \bar{x}^{j_2}, \ldots, \bar{x}^{j_m}]$, a given matrix, and $e = (1, 1, \ldots, 1)$.

I. LP3: $\text{Min } Z(y) = c^t\hat{x} + d^t y + \hat{x}^t C y$

subject to $y \in Y_0$

with $\hat{x}$ fixed.

et $x^1, \ldots, x^m$ be the adjacent extreme points of 0 in X_0. Let $k_0 = \underset{y \in Y_0}{\text{Min}} \{\text{Min}\, c^t x^i + d^t y + (x^i)^t C y$, $\ldots, m\}$. Let $\bar{x}^i$ be the projection of the point x^i obtained by solving the problem LP1 with and $k = k_0$. If $\bar{\lambda}_i$ is the optimal solution to the problem, then $\bar{x}^i = \bar{\lambda}_i x^i$, $i = 1, \ldots, m$. It is easily hat $\bar{\lambda}_i \geq 1$. Let K be the cone with vertex at 0 and whose generators are the rays passing through $\ldots, \bar{x}^m$. We observe that $X_0 \subset K$. The initial polytope P_0 in the sequence is defined by $P_0 = \text{conv}$

$[0, \bar{x}^1, \ldots, \bar{x}^m]$. Its $(m+1)$ facets are $\theta_0, \theta_1, \ldots, \theta_m$, where $\theta_0 = \text{aff}\,[\text{conv}\,[\bar{x}^1, \ldots,$ and $\theta_j = \text{aff}\,[\text{conv}\,[x^1, \ldots, x^{j-1}, 0, x^{j+1}, \ldots, x^m]]$, $j=1, \ldots, m$. Let $L_0 = \{\theta_0\}$ and $S_0 = \{\theta_1, \ldots$

ALGORITHM:

(1) At the ith stage, if $L_i = \phi$, terminate. Otherwise remove an element $\theta_k = \text{aff}\,[\text{conv}\,[\bar{x}^{k_1}, x^{k_m}]]$ from $L_i : L_{i+1} = L_i - \{\theta_k\}$. Solve problem LP2 with B_j formed by the m elements of θ_i $x^q \epsilon X_0$ be a solution of LP2, and let the optimal value be $\bar{Z}$.

(2) If $\bar{Z} = 1$, then $S_{i+1} = S_i \cup \{\theta_k\}$ and go to (1). Otherwise go to (3).

(3) Solve LP3 with $\hat{x} = x^q$. If the optimal value of LP3 is

$\bar{Z}(y) < k_i$, set $\bar{x}^q = x^q$, $k_{i+1} = \bar{Z}(y)$ and go to (4). If
$\bar{Z}(y) = k_i$, set $\bar{x}^q = x^q$, $k_{i+1} = k_i$ and go to (5). If
$\bar{Z}(y) > k_i$, set $k_{i+1} = k_i$ and solve LP1 with $\hat{x} = x^q$ and
$k = k_{i+1}$. Let $\bar{x}^q = \bar{\lambda} x^q$, where $\bar{\lambda}$ is the optimal solution to LP1. Go to (5).

(4) For each extreme point $\bar{x}^j$ of P_i, find the projected point $\bar{\bar{x}}^j$ by solving LP1 with $\hat{x} = \bar{x}$ $k = k_{i+1}$. The polytope P_i and all the facets are defined with respect to these new points $\bar{\bar{x}}^j$. Go to

(5) Express $\bar{x}^q$ as a linear combination of the elements of each $\theta_j \epsilon L_i \cup S_i : \bar{x}^q = B_j \lambda^j$ or $B_j^{-1} \bar{x}^q$, with $B_j = [\bar{x}^j{}_1, \ldots, \bar{x}^j{}_m]$. Find $e\lambda^j$.

(6) Classify each θ_j (including θ_k) into five mutually exclusive sets:

(a) $L_1 = \{\theta_j \epsilon L_i | e\lambda^j < 1\}$,
(b) $L_2 = \{\theta_j \epsilon L_i | e\lambda^j > 1\}$,
(c) $L_3 = \{\theta_j \epsilon L_i | e\lambda^j = 1\}$,
(d) $L_4 = \{\theta_j \epsilon S_i | e\lambda^j < 1\}$,
(e) $L_5 = \{\theta_j \epsilon S_i | e\lambda^j = 1\}$.

(7) Find all pairs θ_m and θ_n with $\theta_m \epsilon L_1 \cup L_4$ and $\theta_n \epsilon L_2$ which have $(m-1)$ elements in com

$\bar{x}^m{}_i = \bar{x}^n{}_i, i = 1, \ldots, (m-1)$. Define
$L_6 = \{(\theta_m, \theta_n) | \theta_m \epsilon L_1 \cup L_4, \theta_n \epsilon L_2, \theta_m$ and θ_n have $(m-1)$ elements in common$\}$
$L_7 = \{\theta_e | \theta_e = \text{aff}\,[\text{conv}\,[\theta_m \cap \theta_n, \bar{x}^q]], (\theta_m, \theta_n) \epsilon L_6\}$,
$L_8 = \{\theta_d | \theta_d = \text{aff}\,[\text{conv}\,[\theta_f, \bar{x}^q]], \theta_f \epsilon L_3\}$,
$L_9 = \{\theta_c | \theta_c = \text{aff}\,[\text{conv}\,[\theta_f, \bar{x}^q]], \theta_f \epsilon L_5\}$.

Then $L_{i+1} = L_1 \cup L_7 \cup L_8$, $S_{i+1} = L_4 \cup L_9$, $P_{i+1} = \text{conv}\,[P_i \cup \{\bar{x}^q\}]$. Go to (1).

Steps (3) and (4) of the algorithm enlarge the polytope P_i and help improve convergence. For polytope P_i in the sequence, its set of facets is $S_i \cup L_i$. Problem LP2 need not be solved in Ste corresponding to any facet in S_i, since we already know that the optimal value will be 1. In Steps (6), and (7), the facets of the polytope P_{i+1} are generated.

Finiteness of the procedure is easily established. In each stage of the algorithm, one of the follo occurs: (i) an extreme point of X_0 not contained in P_i (and hence not contained in any of the polyt generated earlier) is located, or (ii) an element from the finite set L_i is removed. Since both can hap only a finite number of times, the algorithm is finite.

nplementation of this algorithm for realistic problems may require a considerable amount of ıter memory and the capability for efficient list processing. A list corresponding to the sets S_i will have to be stored defining the current set of facets θ_j and the points contained in each facet. he facets have been classified into the sets $L_1, \ldots, L_5$, the sets L_6, L_7, L_8, and L_9 have to be ıined by list processing. As far as computational burden goes, problem LP1 is a parametric linear ming problem. It can be solved by conducting a Bolzano Search for λ over an appropriate interval with a specified tolerance level, where L is a given large number (see [10]). This would require a sequence of linear programs. Problem LP3 is a straightforward linear programing problem. ıain difficulty would come in problem LP2, where we need the inverse of a matrix before the m can be solved. The number of inverses that will need to be evaluated before the procedure ates depends on when the set L_i becomes empty. One way of reducing the number of inverses calculated is to store an appropriate set of inverses, which can be used to generate a much larger inverses. Note that the procedure requires inverses of many matrices which differ from a given in precisely one column. Obviously, there is a tradeoff between memory requirements and the ıtational burden.

ertain practical problems of the form BLP have a special structure. In a location-allocation prob-, the set X_0 has a block diagonal structure with coupling variables, and the set Y_0 has the structure ansportation problem. It is expected that, by applying Rosen's partitioning procedure [6] to the program in X_0, the above algorithm will be applicable to large problems. The basic ideas can be o solve the class of problems considered by Shachtman [7] and Zwart [11].

REFERENCES

alas, E., "Intersection Cuts—A New Type of Cutting Plane for Integer Programming," Operations Research *19*, 19–39 (1971).

ooper, L., "Location-Allocation Problems," Operations Research *11*, 331–343 (1973).

allo, G. and A. Ülkücü, "Bilinear Programming: An Exact Algorithm," Operations Research Center Report No. *ORC 73–26*, University of California, Berkeley, California (1973).

rünbaum, B., *Convex Polytopes* (John Wiley, 1967).

onno, H., "Bilinear Programming," Parts I and II. Technical Report No. *71–9* and *71–10*, Operations Research House, Stanford University (1971).

asdon, L., *Optimization Theory for Large Systems* (Macmillan, 1970).

achtman, R., "Generation of the Admissible Boundary of a Convex Polytope," Operations Research *22*, 151–159 (1974).

achtman, R., Private Communication (May, 1974).

ui, H., *Concave Programming Under Linear Constraints* (Russian), Doklady Akademii Nauk SSSR (1964). English Translation in Soviet Mathematics *5*, 1437–1440 (1964).

aish, H., "Nonconvex Programming, With Applications to Production and Location Problems," Unpublished Ph. D. Dissertation, Georgia Institute of Technology (Dec. 1974).

wart, P. B., "Global Maximization of a Convex Function with Linear Inequality Constraints," Operations Research *22*, 602–609 (1974).

wart, P. B., "Nonlinear Programming: Counter-examples to Two Global Optimization Algorithms," Operations Research *21*, 1260–1266 (1973).

NASH SUBSETS AND MOBILITY CHAINS IN BIMATRIX GAMES

G. A. Heuer*

Concordia College
Moorhead, Minnesota

and

C. B. Millham

Washington State University
Pullman, Washington

ABSTRACT

This work is concerned with constructing, analyzing, and finding "mobility chains" for bimatrix games, sequences of equilibrium points along which it is possible for the two players to progress, one equilibrium point at a time, to an equilibrium point that is preferred by both players. The relationship between mobility chains and Nash subsets is established, and some properties of maximal Nash subsets are proved.

TRODUCTION

bimatrix game is defined by a pair (A, B) of real $m \times n$ matrices together with the Cartesian ct $\mathscr{X} \times \mathscr{Y}$ of all m-dimensional probability vectors $\mathscr{X}$ and all n-dimensional probability vectors $\mathscr{Y}$. nt (x', y') in $\mathscr{X} \times \mathscr{Y}$ is an equilibrium point of the game (A, B) if $x' A y' \geq x A y'$ for all $x \in \mathscr{X}$ and $' \geq x' B y$ for all $y \in \mathscr{Y}$ (transposes assumed where appropriate).

revious work on such games has included the following: Kuhn (5), in simplifying the work of jev (18), defined an extreme equilibrium strategy in a particular way and showed that all extreme brium strategies can be obtained by examining square submatrices of the payoff matrices. Mills owed that a pair (x, y) of mixed strategies with components ξ_i and η_j respectively is an equilib- oint if and only if there exist scalars α, β such that (x, y, α, β) solves: maximize $x A y + x B y$ β subject to

$$Ay \leq \alpha e,\ xB \leq \beta e,\ x \geq \phi_m,\ y \geq \phi_n,\ \sum_{i=1}^{m} \xi_i = \sum_{j=1}^{n} \eta_j = 1,$$

ϕ_k is the k-dimensional 0-vector, appropriately row or column, and e is a row or column vector of whose entries is 1. Mangasarian (8) considered the same problem and pointed to the use of gorithm of Balinski (1) for finding all vertices of a polyhedral convex set in solving it. Lemke and on (6) showed, by an algebraic argument, that an equilibrium point lies on a path joining a sequence acent extreme points of a certain convex polyhedron, and gave an algorithm which terminates

he work of this author was done while visiting at the Mathematisches Institut Der Universität Zu Köln, Germany.

either in an equilibrium point or in an unbounded edge of the polyhedral set. Raghavan (16) studie the properties of equilibrium points in completely mixed games.

The algebraic properties of 0-sum matrix games were explored thoroughly by Gale and Sherma (3) and by Bohnenblust, Karlin, and Shapley (2); some of these properties were shown in (11) to ger eralize to Nash subsets of bimatrix games. Construction of a bimatrix game with given equilibriu points was considered in (10), which also contained a partial result on the old problem of Nash solv bility (7) in bimatrix games. Further results on the problem posed in (10) of constructing a bimatri game having a pair (x', y') as its unique equilibrium point appeared in (4).

There is also an extensive literature on cooperation in finite, two person nonzero sum games datin back to von Neumann and Morgenstern (17) and Nash (13), with later work relating to the solutio of the "Nash bargaining game." (See, e.g., Owen (14) and Pugh and Mayberry (15).)

Between strict noncooperation and cooperation in which players select "cooperative mixed strat gies" (15) leading to a point on the so-called "bargaining line," many situations of economic, politic and perhaps even battleground conflict would appear amenable to analysis through "noncooperativ games with mobility." In this type of analysis, an equilibrium point in the classical sense is attaine perhaps through the algorithm of Lemke and Howson (6); an investigation is then undertaken to se if it is possible by appropriate changes of strategy, first by one player and then the other (never agai being out of equilibrium), to move to another equilibrium point that will be preferred by both player

A Nash subset of a game (A, B) is a set S of equilibrium points which are interchangeable i the sense that if (x^1, y^1) and (x^2, y^2) are in S, then so are (x^1, y^2) and (x^2, y^1). A game is Nash solvabl if all of its equilibrium points lie in a single Nash subset. As mentioned, some properties of Nash sub sets and Nash solvability have been obtained in (10) and (11). For some general discussion of Nas solvability, see (7), pp. 106 ff.

It is clear that mobility requires the existence of nontrivial Nash subsets (i.e., Nash subsets cor taining more than one point). The present work is concerned with recognizing and describing game with mobility and exploring some of the properties of Nash subsets and interchangeability of equilib rium points relevant thereto.

2. EXISTENCE OF MOBILITY CHAINS

If (x, y) is an equilibrium point for (A, B), the expected payoffs to the two players ar $\alpha = x\ A\ y$ and $\beta = x\ B\ y$. (Vectors are regarded as row matrices or column matrices as the contex requires.) At times we shall also refer to the quadruplet $(x, y; \alpha, \beta)$ as an equilibrium point.

DEFINITION 2.1: *A mobility chain of length k* for the bimatrix game (A, B) is a sequence $(x^1$ $y^1; \alpha^1, \beta^1), (x^2, y^1; \alpha^1, \beta^2), (x^2, y^2; \alpha^2, \beta^2), \ldots, (x^k, y^k; \alpha^k, \beta^k)$ of equilibrium points. If $\alpha^i \leq \alpha^{i+}$ and $\beta^i \leq \beta^{i+1}$ for $i = 1, \ldots, k-1$, then the chain is said to be *upwardly mobile*. If $\alpha^i < \alpha^{i+1}, \beta^i < \beta^{i+1}$ the upward mobility is *strict*.

The implication is that while, say, Player 1 has nothing immediate to gain by moving from x^1 t x^2 so long as Player 2 continues to play y^1, he can by doing so enhance the gains of Player 2 at no cos to himself, and Player 2 may in turn reciprocate by improving Player 1's gain without reducing hi own. We direct our attention first to constructing a bimatrix game with a given mobility chain.

We shall denote the ith row of A by $A_{i\cdot}$ and the jth column by $A_{\cdot j}$. If $(x, y; \alpha, \beta)$ is an equilibriun point for the $m \times n$ game (A, B), and $x = (\xi_1, \ldots, \xi_m)$ $y = (\eta_1, \ldots, \eta_n)$, let

$$M_1(x) = \{i : \xi_i > 0\},$$
$$N_1(y) = \{j : \eta_j > 0\},$$
$$M_2(A, y) = \{i : A_i \cdot y = \alpha = \max_r A_r \cdot y\},$$
$$N_2(x, B) = \{j : x B \cdot_j = \beta = \max_r x B \cdot_r\}.$$

f u, w are both n-dimensional vectors, let $< u, w >$ be the inner product of u and w.
ne following is an easy consequence of the definitions. (See [10], Theorem 4.)

EMARK 2.2: Necessary and sufficient conditions for (x, y) to be an equilibrium point for the $A, B)$ are that $M_1(x) \subseteq M_2(A, y)$ and $N_1(y) \subseteq N_2(x, B)$.

HEOREM 2.3: Let $\{x^1, \ldots, x^k\}, \{y^1, \ldots, y^k\}$ be given sets of probability vectors in E^m, E^n, tively. Let A and B be real $m \times n$ matrices, and for

$$1 \leqslant r \leqslant k, \text{ let } \alpha^r = \max_i A_i \cdot y^r, \beta^r = \max_j x^r B \cdot_j.$$

ssary and sufficient condition for

$$; \alpha^1, \beta^1), (x^2, y^1; \alpha^1, \beta^2), (x^2, y^2; \alpha^2, \beta^2), \ldots, (x^k, y^{k-1}; \alpha^{k-1}, \beta^k), (x^k, y^k; \alpha^k, \beta^k)$$

mobility chain for (A, B) is that

$$M_1(x^i) \subseteq M_2(A, y^j) \quad \text{for } j = i, i-1;\ 2 \leqq i \leqq k;$$
$$M_1(x^1) \subseteq M_2(A, y^1);$$
$$N_1(y^j) \subseteq N_2(x^i, B) \quad \text{for } i = j, j+1;\ 1 \leqq j \leqq k-1; \text{ and}$$
$$N_1(y^k) \subseteq N_2(x^k, B).$$

ROOF: This follows immediately from (2.2).

HEOREM 2.4: Let $X = \{x^1, \ldots, x^k\}$ and $Y = \{y^1, \ldots, y^k\}$ be given linearly independent probability vectors in E^m, E^n respectively. Let $\alpha^1, \ldots, \alpha^k$ and $\beta^1, \ldots, \beta^k$ be given positive umbers. There exists a game (A, B) for which $(x^1, y^1; \alpha^1, \beta^1), (x^2, y^1; \alpha^1, \beta^2), (x^2, y^2; \alpha^2, \ldots, (x^k, y^k; \alpha^k, \beta^k)$ is a mobility chain.

ROOF: For each $i \in \cup_{x \in X} M_1(x)$, let

$$X(i) = \{x \in X : \xi_i > 0\};$$
$$S_i = \{j : (x, y^j) \text{ is to be an equilibrium point for some } x \text{ in } X(i)\}.$$

or $1 \leqq j \leqq k$, choose a^j in E^n such that $< a^j, y^j > = \alpha^j$ and $< a^j, y^i > = 0$ for $i \neq j$. Let

$$A_{i\cdot} = \sum_{j=1}^{k} \gamma_{ij} a^j + \sum_{j=1}^{n-k} \lambda_{ij} \bar{a}^j,$$

$\{\bar{a}^1, \ldots, \bar{a}^{n-k}\}$ is a basis for the null space of the matrix with rows $y^1, \ldots, y^k$ (regarded ear transformation from E^n to E^k), $\gamma_{ij} = 1$ if $j \in S_i$, $\gamma_{ij} \leqq 1$ if $j \notin S_i$, and λ_{ij} is arbitrary. Then

$$A_{i\cdot}\, y^j = \langle \gamma_{ij}\, a^j,\, y^j \rangle = \begin{cases} \alpha^j & , \text{ if } j \in S_i \\ \gamma_{ij}\, \alpha^j \leqq \alpha^j , & \text{if } j \notin S_i . \end{cases}$$

It follows that for each (x, y) that was to be an equilibrium point, $M_1(x) \subseteq M_2(A, y)$. For, if ξ_i then $x \in X(i)$, $y = y^j \in S_i$ for some j, and $A_{i\cdot} y = \alpha^j \geqq A_{r\cdot} y$ for each r, $1 \leqq r \leqq m$.

After a similar construction of B, we have a game (A, B) with the desired mobility chain.

Note that we have shown more than was demanded; with X and Y as given, we may const (A, B) so that any desired subset of $X \times Y$ consists entirely of equilibrium points.

3. MAXIMAL NASH SUBSETS

It is clear that each successive pair of points in a mobility chain has the Nash interchangeab property. It will occur to some readers to ask whether it is not possible in general to proceed dir $(x^1, y^1; \alpha^1, \beta^1)$, $(x^k, y^1; \alpha^1, \beta^k)$, $(x^k, y^k; \alpha^k, \beta^k)$. That the answer is negative, i.e., that there mobility chains of irreducible length $k > 2$, depends on the fact that the relation of Nash intercha ability between equilibrium points is not transitive. Consequently, two distinct "maximal" (D tion 3.1) Nash subsets may overlap, and to move from a point of one to a point of the o without ever being out of equilibrium is possible by first moving to a point of the intersection. In section we investigate to some extent the structure of maximal Nash subsets and show that when of them intersect, they have an extreme point in common. This fact is useful in searching for mobility chains beginning at a given equilibrium point of a given game.

If (x^1, y^1) and (x^2, y^2) are equilibrium points for a given bimatrix game (A, B), we shall write $y^1) \sim (x^2, y^2)$ to mean that (x^1, y^1) and (x^2, y^2) are Nash-interchangeable. Let $I(x^\circ, y^\circ) = \{|(x, y)$ $y) \sim (x^\circ, y^\circ)\}$, the set of all equilibrium points which are interchangeable with (x°, y°).

DEFINITION 3.1: *A maximal Nash subset* is a Nash subset which is not properly contai in any Nash subset.

Clearly $I(x, y)$ contains any Nash subset containing (x, y), but, as we shall see, it need itself be a Nash subset.

LEMMA 3.2: If (x^1, y) and (x^2, y) are equilibrium points for a given bimatrix game (A, B), (x, y) for every convex combination $x = \lambda x^1 + (1-\lambda)x^2$, $0 \leqq \lambda \leqq 1$. A similar statement holds (x, y^1) and (x, y^2).

The proof may be made easily from (2.2), and we omit the details.

THEOREM 3.3: Maximal Nash subsets are convex and closed, as subsets of $E^m \times E^n$.

PROOF: The convexity follows from the preceding lemma, though not quite as immediatel one might first suspect. Let (x^1, y^1) and (x^2, y^2) be elements of a maximal Nash subset S, and $(x, y$ $\lambda(x^1, y^1) + (1-\lambda)(x^2, y^2)$, $0 \leqq \lambda \leqq 1$. Then (x^2, y^1) and (x^1, y^2) are also in S, and by three succes applications of the lemma, we have that (x, y^1), (x, y^2) and then (x, y) is an equilibrium point. To s (x, y) is in S, let (u, v) be any point of S. Then (x^1, v) and (x^2, v) are equilibrium points, so, by the lem (x, v) is; similarly (u, y) is an equilibrium point, so (x, y) is Nash-interchangeable with each p of S.

To show S is closed, let (x, y) be a limit point of S and $\{(x^r, y^r)\}$ be a sequence of points in S verging to (x, y). We must show (x, y) is an equilibrium point interchangeable with every point in S. We use (2.2). If $j \in M_1(x)$, then $j \in M_1(x^r)$ for large r; i.e., $M_1(x) \subseteq M_1(x^r)$ for large r. Since

an equilibrium point for every s, $M_1(x^r) \subseteq \bigcap_{s=1}^{\infty} M_2(A, y^s)$. Now let $i \in \bigcap_{s=1}^{\infty} M_2(A, y^s)$. Then for , $A_{i.}y^s = \max_k A_{k.}y^s$. If $A_{i.}y < A_{k.}y$ for some k, then by the continuity of the product, $A_{i.}y^s < A_{k.}y^s$ ne s. Thus $A_{i.}y = \max_k A_{k.}y$: i.e., $i \in M_2(A, y)$. We have therefore, for a suitably large r $M_1(x) \subseteq$ $) \subseteq \bigcap_{s=1}^{\infty} M_2(A, y^s) \subseteq M_2(A, y)$. By a similar argument, we see that $N_1(y) \subseteq N_2(x, B)$, so (x, y) equilibrium point.

uppose now that $(u, v) \in S$. Then for each s, (u, y^s) is an equilibrium point, so

$$M_1(u) \subseteq \bigcap_{s=1}^{\infty} M_2(A, y^s) \subseteq M_2(A, y).$$

or each r, $M_1(x^r) \subseteq M_2(A, v)$ because $(x^r, v) \in S$; so for a suitably large r, $M_1(x) \subseteq M_1(x^r) \subseteq$ $v)$. Similarly, $N_1(v) \subseteq N_2(x, B)$ and $N_1(y) \subseteq N_2(u, B)$, so (x, v) and (u, y) are equilibrium Thus S is closed.

HEOREM 3.4:

. The relation $\sim$ in the set of all equilibrium points of a bimatrix game is reflexive and sym-, but not necessarily transitive.

Distinct maximal Nash subsets need not be disjoint.

. The set $I(x^\circ, y^\circ)$ may fail to be a Nash subset.

. $I(x^\circ, y^\circ)$ may fail to be convex.

ROOF: That $\sim$ is reflexive and symmetric is obvious. The remaining assertions may be demon- by an example. Let

$$A = \begin{bmatrix} 9 & 9 & 9 \\ 9 & 8 & 11 \\ 8 & 7 & 16 \\ 10 & 8 & 8 \end{bmatrix}, \qquad B = \begin{bmatrix} 2 & 2 & 2 \\ 3 & 3 & 3 \\ 4 & 1 & 3 \\ 1 & 4 & 2 \end{bmatrix}.$$

aximal Nash subsets for (A, B) may be shown to be:

$$S_1 = \{(x, y) : x = (1, 0, 0, 0),\ 0 \leqq \eta_1 \leqq 1/2,\ \eta_2 \geqq (7 - 8\eta_1)/9\},$$
$$S_2 = \{(x, y) : \xi_3 = \xi_4,\ y = (1/2, 1/3, 1/6)\},$$
$$S_3 = \{(x, y) : x = (0, 0, 1/2, 1/2),\ 1/2 \leqq \eta_1 \leqq 4/5,\ \eta_2 = (8 - 10\eta_1)/9\}.$$

derstood, of course, that x and y are always probability vectors.

verify that the points of $S_1 \cup S_2 \cup S_3$ are equilibrium points is straightforward. Moreover, each osed under interchange; i.e., if (x^1, y^1) and (x^2, y^2) are in S_i, so are (x^1, y^2) and (x^2, y^1). Thus are Nash subsets. It will be clear that they are maximal Nash subsets if it is shown that no further rium points exist. While we leave the details of this verification to the reader, the following out- ay be helpful.

Suppose that (x, y) is an equilibrium point. If $\xi_1\xi_2 \neq 0$ and $\xi_3 = \xi_4 = 0$, one finds that $y =$ $1/3, 1/6)$, so that $(x, y) \epsilon S_2$. If $\xi_1 = 1$, one shows $(x, y) \epsilon S_1$; and if $\xi_2 = 1$, $y = (1/2, 1/3, 1/6)$ an $y) \epsilon S_2$. This takes care of all cases with $\xi_3 = \xi_4 = 0$. If $\xi_3 \neq 0$ and $\xi_4 = 0$, one finds y must be $(1, ($ this would imply $\xi_3 = 0$, so this case cannot occur. Similarly $\xi_3 = 0$, $\xi_4 \neq 0$ implies $y = (0, 1, 0)$, ν requires $\xi_4 = 0$; again this is impossible. In all remaining cases $\xi_3\xi_4 \neq 0$. Here one finds $1/2 \leqq \eta_1 \leqq$ $\eta_2 = (8 - 10\eta_1)/9$ and $\eta_3 = (1 + \eta_1)/9$. Thus $\eta_1\eta_3 > 0$, and this leads to $\xi_3 = \xi_4$. There are two cases: If $\xi_1 \neq 0$ or $\xi_2 \neq 0$, we must have $y = (1/2, 1/3, 1/6)$ and $(x, y) \epsilon S_2$. If $\xi_1 = \xi_2 = 0$, $(x, y$

Note that $S_1 \cap S_2 = ((1, 0, 0, 0), (1/2, 1/3, 1/6))$ and $S_2 \cap S_3 = ((0, 0, 1/2, 1/2), (1/2, 1/3,$ This verifies (A) and (B). $I((1, 0, 0, 0), (1/2, 1/3, 1/6)) = S_1 \cup S_2$, which is not convex and hence Nash subset. This verifies (C) and (D).

Beginning at the point $((1, 0, 0, 0), (0, 7/9, 2/9))$ in S_1, there is an irreducibly mobility chain term ing in S_3 illustrated in the lower portion of the following diagram. Near the top of each colum general values of α and β for the Nash subset in that column.

S_1	$S_1 \cap S_2$	S_2	$S_2 \cap S_3$	S_3
$\alpha = 9$ $\beta = 2$		$\alpha = 9$ $\beta = 2 + \xi_2 + \xi_3$		$\alpha = 8 + 2\eta_1$ $\beta = 2.5$
$x^1 = (1, 0, 0, 0)$ $y^1 = (0, 7/9, 2/9)$ $\alpha^1 = 9$ $\beta^1 = 2$	$x^1 = (1, 0, 0, 0)$ $y^2 = (1/2, 1/3, 1/6)$ $\alpha^2 = 9$ $\beta^1 = 2$		$x^2 = (0, 0, 1/2, 1/2)$ $y^2 = (1/2, 1/3, 1/6)$ $\alpha^2 = 9$ $\beta^2 = 2.5$	$x^2 = (0, 0, 1/2, 1/$ $y^3 = (4/5, 0, 1/5)$ $\alpha^3 = 9.6$ $\beta^2 = 2.5$

(Actually, we have Player 2 changing strategies first, but of course the roles of the two players cou reversed by making obvious changes in the example.) Note that while the first step did not imp either player's gain, it was a necessary step for the improvement ultimately achieved by both.

The above example is interesting for another reason. Although this is not a Nash-solvable gan is possible to move from any one equilibrium point to any other via a mobility chain, i.e., wit ever being out of equilibrium.

LEMMA 3.5:

(A) Suppose x^1 is in equilibrium with y^1 and with y^2, and x^2 is in equilibrium with a point $\bar{y} = \lambda$ $(1 - \lambda) y^2$ for some λ, $0 < \lambda < 1$. Then x^2 is in equilibrium with y^1 and y^2 also.

(B) A similar result holds with the roles of x and y interchanged.

PROOF:

(A) Note first that

$$N_1(y^1) \cup N_1(y^2) \subseteq N_1(\bar{y}) \subseteq N_2(x^2, B), \tag{3.5.1}$$

and next that $\emptyset \neq M_1(x^1) \subseteq M_2(A, y^1) \cap M_2(A, y^2)$. For $1 \leqq i \leqq m$, $A_i.\bar{y} = \lambda A_i.y^1 + (1 - \lambda)A$ Since both λ and $1 - \lambda$ are positive and $M_2(A, y^1) \cap M_2(A, y^2) \neq \emptyset$, it follows that $M_2(A, \bar{y}) = M_2$ $y^1) \cap M_2(A, y^2)$. Thus

$$M_1(x^2) \subseteq M_2(A, \bar{y}) = M_2(A, y^1) \cap M_2(A, y^2). \tag{3.5.2}$$

rom (3.5.1) and (3.5.2), the desired result follows by an appeal to (2.2).

3) is proved similarly.

HEOREM 3.6: Let S_1 and S_2 be maximal Nash subsets for a game (A, B), and $\bar{p} = (\bar{u}, \bar{v}) \in S_1 \cap$
en $\bar{p}$ is an extreme point of both S_1 and S_2 or of neither.

lore specifically, if $\bar{p} = \lambda p^1 + (1-\lambda)p^2$ for some p^1 and p^2 in S_1, with $p^1 \neq p^2$, $0 < \lambda < 1$, then
p^2 are in S_2 also.

ROOF: Let $q^1 = (x^1, y^1) \in S_1$, $q^2 = (x^2, y^2) \in S_2$, and suppose $\bar{p} = \lambda p^1 + (1-\lambda)p^2$ for $p^i = (u^i, v^i)$
$i = 1, 2$. Since $q^1 \sim p^1$ and $q^1 \sim p^2$, x^1 is in equilibrium with v^1 and v^2. Similarly, since $\bar{p}$ is in
is in equilibrium with $\bar{v} = \lambda v^1 + (1-\lambda)v^2$. By Lemma 3.5, x^2 is in equilibrium with v^1 and v^2.
y a similar argument, one sees that y^2 is in equilibrium with u^1 and u^2. Thus $p^1 \sim q^2$ and $p^2 \sim q^2$.
this holds for all q^2 in S_2, we have p^1 and p^2 in S_2. This proves the theorem.

OROLLARY 3.7: Every extreme point of $S_1 \cap S_2$ is an extreme point of S_1 and of S_2.

ROOF: Let $\bar{p}$ be an extreme point of $S_1 \cap S_2$, $q^1 \in S_1$, $q^2 \in S_2$. If $\bar{p}$ is not an extreme point of S_1,
$= (p^1 + p^2)/2$ for some $p^1, p^2 \in S_1$, $p^1 \neq p^2$. Then by the theorem, p^1 and p^2 are in S_2, so they are
$\cap S_2$, contrary to the fact that $\bar{p}$ is an extreme point of $S_1 \cap S_2$.

OROLLARY 3.8: If S_1 and S_2 are distinct maximal Nash subsets for a game (A, B), and $S_1 \cap$
, then $S_1 \cap S_2$ contains a point which is an extreme point of S_1 and of S_2.

ROOF: Since $S_1 \cap S_2$ is a nonempty bounded closed convex set, it has extreme points. The result
llows from Corollary 3.7.

OCATION OF MOBILITY CHAINS

ne natural way to search for mobility chains is be linear programming. The process is initiated at
n equilibrium point, which can be found using complementary pivot theory. (It is to be noted that
e with mobility is degenerate and complementary pivot theory assumes nondegeneracy, but that
legenerate game can be derived by perturbation from a degenerate game and an equilibrium point
nondegenerate game found. For sufficiently small perturbations, this point will be an equilibrium
of the degenerate game.) Let (x^1, y^1) be such an equilibrium point, with payoffs (α^1, β^1), and
he following linear programing problem:

$$
\begin{cases}
\textit{Maximize } \beta \textit{ subject to:} & \\
xB_{\cdot j} = \beta & \textit{for } j \in N_1(y^1); \\
xB_{\cdot j} \leqq \beta & \textit{for } j \notin N_1(y^1),\ 1 \leqq j \leqq n; \\
\xi_i = 0 & \textit{for } i \notin M_2(A, y^1)\ 1 \leqq i \leqq m; \\
\xi_i \geqq 0, \quad \sum_{i=1}^{m} \xi_i = 1. &
\end{cases}
$$

the program has a solution x^2, with $\beta^2 \neq \beta^1$, then $(x^2, y^1; \alpha^1, \beta^2)$ is an equilibrium point for
upwardly mobile (strictly) from $(x^1, y^1; \alpha^1, \beta^1)$. If the solution is $x^2 \neq x^1$ but $\beta^2 = \beta^1$, upward
ty but not strict upward mobility exists.

corresponding solution to

$$(4.2)\quad \begin{cases} \textit{Maximize } \alpha \textit{ subject to:} \\ \quad A_{i\cdot}y = \alpha \quad \textit{for } i \in M_1(x^2); \\ \quad A_{i\cdot}y \leqq \alpha \quad \textit{for } i \notin M_1(x^2),\ 1 \leqq i \leqq m; \\ \quad \eta_j = 0 \quad \textit{for } j \notin N_2(x^2, B),\ 1 \leqq j \leqq n; \\ \quad \eta_j \geqq 0, \quad \sum_{j=1}^{n} \eta_j = 1 \end{cases}$$

will determine whether there is upward mobility from $(x^2, y^1; \alpha^1, \beta^2)$ to a point $(x^2, y^2; \alpha^2, \beta^2)$.

If (x^1, y^1) lies in a maximal Nash subset S_1 which intersects no other maximal Nash subse the mobility chain must terminate here. (The final point is Nash-interchangeable with the initial o so it is attainable via one intermediate point.) On the other hand, if there are other maximal N subsets intersecting S_1, we know from Corollary 3.8 that at least one extreme point of S_1 lies in intersection. Suppose that x^2 is a solution to (4.1). We may assume that x^2 is an extreme point of constraint set. From the definition of maximal Nash subset, we see that $S_1 = X_1 \times Y_1$, where $X_1 = \{x:$ $y) \in S_1$ for some $y\}$ and $Y_1 = \{y: (x, y) \in S_1$ for some $x\}$.

We claim that x^2 is an extreme point of X_1. For if it were not, then $x^2 = (u^1 + u^2)/2$ for s $u^1 \neq u^2$ in X_1. But since (u^1, y^1) and (u^2, y^1) are equilibrium points, u^1 and u^2 are in the constr set, contrary to the assumption that x^2 was an extreme point of the constraint set.

If y^2 is a solution to (4.2), also occurring at an extreme point of the constraint set, then y^2 i extreme point of Y_1, and therefore (x^2, y^2) is an extreme point of S_1.

There is no guarantee, unfortunately, that (x^2, y^2) is one of the extreme points which lie in intersection of S_1 with another maximal Nash subset. Indeed, in the game given in the proof of Th rem 3.4, we have a counterexample: In S_2, α is constant, and the maximum value of β occurs w $x = (0, 1, 0, 0)$ (giving $\beta = 3$) rather than at the point $x = (0, 0, 1/2, 1/2)$ which admits us to S_3.

Nonetheless, the way to the next maximal Nash subset is through an extreme point of the pre one, and when the chain formed by successive maximizing of α and β terminates, it may be poss to extend the mobility chain further by seeking out other extreme points of the constraint set at var stages.

REFERENCES

[1] Balinski, M. L., "An Algorithm for Finding all Vertices of Convex Polyhedral Sets," J. Soc. Ind Appl. Math. *9:* 72–88 (1961).

[2] Bohnenblust, H. F., S. Karlin, and L. S. Shapley, "Solutions of Discrete, Two-Person Gam in *Contributions to the Theory of Games*, Annals of Math. Studies *24*, Princeton University P (1950).

[3] Gale, D. and S. Sherman, "Solutions of Finite Two-Person Games," in *Contributions to Theory of Games*, Annals of Math. Studies *24*, Princeton University Press (1950).

[4] Heuer, G. A., "Uniqueness of Equilibrium Points in Bimatrix Games," to appear, Int. J. G Theory.

[5] Kuhn, H. W., "An Algorithm for Equilibrium Points in Bimatrix Games," Proc. N.A.S. *47:* 1 1662 (1961).

[6] Lemke, C. E. and J. T. Howson, Jr., "Equilibrium Points of Bimatrix Games," J. Soc. In Appl. Math. *12:* 413–423 (1964).

ce, R. D. and H. Raiffa. *Games and Decisions* (John Wiley and Sons, N.Y., 1957).
ngasarian, O. L., "Equilibrium Points of Bimatrix Games," J. Soc. Indust. Appl. Math. *12:* 778–780 (1964).
llham, C. B, "On the Structure of Equilibrium Points in Bimatrix Games," SIAM Review *10:* 447–449 (1968).
—— "Constructing Bimatrix Games with Special Properties," Nav. Res. Log. Quart. *19:* 709–714 (1972).
—— "On Nash Subsets of Bimatrix Games," Nav. Res. Log. Quart. *21:* 307–317 (1974).
lls, H., "Equilibrium Points in Finite Games," J. Soc. Indust. Appl. Math. *8:* 397–402 (1960).
sh, J. F. Jr., "Two-Person Cooperative Games," Econometrica *21:* 128–140 (1953).
ven, G., "Optimal Threat Strategies in Bimatrix Games," Int. J. Game Theory *1:* 3–9 (1971).
gh, G. E. and J. P. Mayberry, "Theory of Measure of Effectiveness for General-Purpose Military Forces, Part I: A Zero-Sum Payoff Appropriate for Evaluating Combat Strategies," Operations Research *21:* 867–885 (1973).
ghavan, T. E. S., "Completely Mixed Strategies in Bimatrix Games," J. London Math. Soc. *2:* 709–712 (1970).
n Neumann, J. and O. Morganstern. *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, N.J., 1953), 3rd. Ed.
robjev, N. N., "Equilibrium Points in Bimatrix Games," Theoriya Veroyatnostej i ee Primenenya *3:* 318–331 (1958).

THE CORE AND COMPETITIVE EQUILIBRIA OF A MARKET WITH INDIVISIBLE GOODS

Mamoru Kaneko

Department of Social Engineering
Tokyo Institute of Technology
Tokyo, Japan

ABSTRACT

We consider a generalization of the assignment game of Shapley and Shubik [4]. In the market which we consider, s kinds of indivisible goods are exchanged for money. The market consists of buyers and sellers. Each buyer wants to buy at most one unit of the goods, and each seller may sell more than one unit. First, we show that the set of all competitive imputations is given by the solutions of a certain linear programing problem dual to the optimal problem. Second, we show that the core of the market coincides with the set of all competitive imputations under some condition, and consider the core of the market where $s=1$ and the condition does not hold.

TRODUCTION

apley and Shubik [4] formulated a simple and elegant game, the assignment game, which is a for two-sided markets in which a product that comes in large, indivisible units is exchanged ney, and in which each participant either supplies or demands exactly one unit. They showed e core of the market game is the solutions of a certain linear programing problem dual to the l assignment problem, and that it coincides with the set of the competitive imputations—the tions given by competitive equilibria—of the market.

this paper we consider a generalization of the market. The market which we consider consists ers and sellers. Each buyer wants to buy at most one unit, but each seller may sell more than it. This market is more applicable to markets where large and indivisible goods (e.g., houses, tc.) are exchanged.

rst, we show that the set of all competitive imputations is given as the solutions of a certain programing problem dual to the optimal problem. Second, we show that the core coincides with of all competitive imputations under some condition, and consider the core of a market with nd of good in the case where the condition does not hold.

E MARKETS (M, N) AND (M^*, N)

the market (M, N) the set of traders consists of M and N, where $M=\{1, \ldots, m\}$ is the set of and $N=\{1', \ldots, n'\}$ the set of buyers. In the market (M, N) s kinds of indivisible goods are ged for money. Each seller $i \in M$ owns $w^i=(w^i_1, \ldots, w^i_s)$ initially, where w^i_k $(k=1, \ldots, s)$ nnegative integer. Each buyer $j \in N$ owns no good initially. Each trader has a utility function ed in terms of money. Let seller i's utility function be $U_i(x^i)$. We assume:

$$U_i(x^i) = U_i(x_1^i, \ldots, x_s^i) = \sum_{k=1}^{s} U_i(x_k^i e^k), \tag{2.1}$$

where e^k is the vector such that $e_k^k = 1$ and $e_t^k = 0$ for $\forall\ t \neq k$. Here we put $U_i(0) = 0$ for $\forall\ i \in M$. assumption (2.1) means that the utility of an amount of one good is independent of amounts of goods. By putting $a_i(k, f) = U_i(fe^k) - U_i((f-1)e^k)$, we can write

$$U_i(x^i) = \sum_{k=1}^{s} \sum_{f=1}^{x_k^i} a_i(k, f). \tag{2.2}$$

We assume that for $\forall k = 1, \ldots, s$:

$$a_i(k, f) \geqq 0 \qquad \text{for } \forall f \geqq 1, \tag{2.3}$$

$$a_i(k, f) \geqq a_i(k, f+1) \qquad \text{for } \forall f \geqq 1, \tag{2.4}$$

$$a_i(k, f) = 0 \qquad \text{for } \forall f > w_k^i. \tag{2.5}$$

Condition (2.3) means monotonicity and (2.4) convexity of $U_i(x^i)$. Condition (2.5) means that the u of seller i does not increase though he owns more than w_k^i units of kth good.

Let buyer j's utility functions be $U_j(x^j)$. We assume:

$$U_j(fe^K) = \begin{cases} h_{kj} \geqq 0 & \text{if } f > 0 \\ 0 & \text{if } f = 0, \end{cases} \tag{2.6}$$

$$U_j(x^j) = \max \{h_{kj} | x_k^j > 0\}. \tag{2.7}$$

Conditions (2.6) and (2.7) mean that the utility of buyer j does not increase though he owns more one unit.

The characteristic function of the market (M, N) is given by

$$v(S) = \max \left(\sum_{i \in S \cap M} U_i(x^i) + \sum_{j \in S \cap N} U_j(x^j) \right) \tag{2.8}$$

subject to

$$\sum_{i \in S \cap M} w^i = \sum_{i \in S \cap M} x^i + \sum_{j \in S \cap N} x^j.$$

In order to analyze the market (M, N), we consider another market (M^*, N). The set of tra in (M^*, N) consists of M^* and N, where N is the same as the set of buyers in (M, N) and M^* is give

$$M^* = \bigcup_{i \in M} \bigcup_{k=1}^{s} \{i(k, 1), \ldots, i(k, w_k^i)\}. \tag{2.9}$$

ility function of seller $i(k,f) \in M^*$ is given by

$$U_{i(k,f)}(x^{i(k,f)}) = \begin{cases} a_i(k,f) & \text{if } x_k^{i(k,f)} > 0 \\ 0 & \text{if } x_k^{i(k,f)} = 0. \end{cases}$$

$i(k,f) \in M^*$ owns one unit of kth good initially. The market (M^*, N) is a special one of (M, N). e define a characteristic function v^* of the market (M^*, N) as follows:

$$v^*(\{t\}) = 0 \quad \text{for } \forall t \in M^* \cup N,$$

$$v^*(\{i(k,f), j\}) = \max(h_{kj} - a_i(k,f), 0) \qquad \text{for } \forall i(k,f) \in M^*, j \in N.$$

t $v^*(\{i(k,f), j\}) = b_{i(k,f)j}$. For a coalition $S \subset M^* \cup N$, we define $v^*(S)$ by

$$v^*(S) = \max(b_{t_1 j_1} + \ldots + b_{t_f j_f})$$

aximum to be taken over all arrangements of $2f$ distinct traders $t_1, \ldots, t_f \in S \cap M^*$ and $\ldots, j_f \in S \cap N$ where $f = \min(|S \cap M^*|, |S \cap N|)$. Here $|S|$ is the number of members in the

he game v^* is an assignment game and is a normalization of the characteristic function given) in the market (M^*, N) such that $v^*(\{t\}) = 0$ for $\forall t \in M^* \cup N$. Equation (2.13) is represented assignment problem as follows:

$$v^*(S) = \max \sum_{t \in S \cap M^*} \sum_{j \in S \cap N} b_{tj} X_{tj}$$

t to

$$\sum_{t \in S \cap M^*} X_{tj} \leqq 1 \qquad \text{for } \forall j \in S \cap N$$

$$\sum_{j \in S \cap N} X_{tj} \leqq 1 \qquad \text{for } \forall t \in S \cap M^*$$

$$X_{tj} \geqq 0 \qquad \text{for } \forall t \in S \cap M^*, j \in S \cap N.$$

he dual problem of the assignment problem (2.14) for the coalition $M^* \cup N$ is

$$\min \left(\sum_{t \in M^*} u_t^* + \sum_{j \in N} v_j^* \right)$$

$$\text{subject to } u_t^* + v_j^* \geqq b_{tj},\ u_t^* \geqq 0,\ v_j^* \geqq 0 \text{ for } \forall t \in M^*, j \in N.$$

be the set of all optimal solutions of (2.15). It is easily verified that (2.14) and (2.15) have optimal ns.

EXAMPLE 1: Let $M=\{1, 2\}$, $N=\{1', 2'\}$, $s=2$, $w_1^1=w_2^2=1$, $w_2^1=w_1^2=2$. Let the utility func be given by the following tables:

TABLE 1

k \ f	1	2
1	5	
2	3	2

$a_1(k, f)$

TABLE 2

k \ f	1	2
1	6	4
2	8	

$a_2(k, f)$

TABLE 3

k \ j	1′	2′
1	10	15
2	15	10

h_{kj}

The characteristic function v is given by $v(\{1'\})=v(\{2'\})=v(\{1', 2'\})=0$, $v(\{1\})=10$, $v(\{2\})=$ $v(\{1, 2\})=28$, $v(\{1, 1'\})=23$, $v(\{1, 2'\})=20$, $v(\{1, 1', 2'\})=33$, $v(\{2, 1'\})=25$, $v(\{2, 2'\})=$ $v(\{2, 1', 2'\})=36$, $v(\{1, 2, 1'\})=41$, $v(\{1, 2, 2'\})=39$, $v(\{1, 2, 1', 2'\})=52$. In this example, A $\{1(1, 1), 1(2, 1), 1(2, 2), 2(1, 1), 2(1, 2), 2(2, 1)\}$. The value of (b_{tj}) is given by the following t

TABLE 4

j \ t	1(1, 1)	1(2, 1)	1(2, 2)	2(1, 1)	2(1, 2)	2(2, 1)
1′	5	12	13	4	6	7
2′	10	7	8	9	11	2

The optimal assignment of (2.14) for $M^* \cup N$ is

$$X_{1(2,2)1'}=X_{2(1,2)2'}=1, X_{tj}=0 \text{ if otherwise.}$$

A^* is given by

$$\{(0, 0, u^*_{1(2,2)}, 0, u^*_{2(1,2)}, 0, 13-u^*_{1(2,2)}, 11-u^*_{2(1,2)}) \mid 0 \leqq u^*_{1(2,2)} \leqq 1, 0 \leqq u^*_{2(1,2)} \leqq 1\}.$$

A payoff vector (u, v) is called an imputation if (u, v) satisfies

$$\sum_{i \in M} u_i + \sum_{j \in N} v_j = v(M \cup N), \tag{2.16}$$

$$u_i \geqq v(\{i\}) \quad \text{for } \forall i \in M, \tag{2.17}$$

$$v_j \geqq v(\{j\}) \quad \text{for } \forall j \in N.$$

The core of the game v is the set of all imputations which satisfy

$$\sum_{i \in S \cap M} u_i + \sum_{j \in S \cap N} v_j \geqq v(S) \qquad \text{for } \forall S \subset M \cup N.$$

$p, x) = (p_1, \ldots, p_s, x^1, \ldots, x^m, x^{1'}, \ldots, x^{n'})$ is called a competitive equilibrium if (p, x)

$$U_i(x^i) + p(w^i - x^i) = \max_z (U_i(z) + p(w^i - z)) \qquad \text{for } \forall i \in M$$

$$U_j(x^j) - px^j = \max_z (U_j(z) - pz) \qquad \text{for } \forall j \in N,$$

$$\sum_{i \in M} x^i + \sum_{j \in N} x^j = \sum_{i \in M} w^i.$$

f vector (u, v) is called a competitive imputation if there is a competitive equilibrium (p, x) at

$$u_i = U_i(x^i) + p(w^i - x^i) \qquad \text{for } \forall i \in M$$

$$v_j = U_j(x^j) - px^j \qquad \text{for } \forall j \in N.$$

he market (M^*, N) the core coincides with A^* and the set of all competitive imputations, which wn by Shapley and Shubik [4].

MMA 1: Each competitive imputation belongs to the core.

OOF: Let (u, v) be a competitive imputation, and let (p, x) be a competitive equilibrium giving Let $v(S) = \sum_{i \in S \cap M} U_i(y^i) + \sum_{j \in S \cap N} U_j(y^j)$. It follows from (2.19) that

$$\sum_{i \in S \cap M} u_i + \sum_{j \in S \cap N} v_j$$

$$= \sum_{i \in S \cap M} (U_i(x^i) + p(w^i - x^i)) + \sum_{j \in S \cap N} (U_j(x^j) - px^j)$$

$$\geqq \sum_{i \in S \cap M} (U_i(y^i) + p(w^i - y^i)) + \sum_{j \in S \cap N} (U_j(y^j) - py^j)$$

$$= \sum_{i \in S \cap M} U_i(y^i) + \sum_{j \in S \cap N} U_j(y^j) = v(S).$$

Q.E.D.

an arbitrary $S \subset M \cup N$, we put

$$S^* = \bigcup_{i \in S \cap M} \bigcup_{k=1}^{s} \{i(k, 1), \ldots, i(k, w_k^i)\} \cup (S \cap N).$$

is easily verified that

$$v(S) = v^*(S) + \sum_{i \in S \cap M} \sum_{k=1}^{s} \sum_{f=1}^{w_k^i} a_i(k, f).$$

For an arbitrary optimal solution X of (2.14) for an S^*, we define Y by

$$Y_{i(k,f)j}=\begin{cases} X_{i(k,f)j} & \text{if } h_{kj} \geqq a_i(k,f) \\ 0 & \text{if } h_{kj} < a_i(k,f). \end{cases} \tag{2.24}$$

LEMMA 2: Let X be an extreme point solution of (2.14) for S^*. For Y given by (2.24), we put

$$T_{ik}=\{j \in S \cap N \mid \sum_{f=1}^{w_k^i} Y_{i(k,f)j}=1\} \text{ for } \forall i \in S \cap M,\ k=1, \ldots, s. \tag{2.25}$$

Then we have

$$v(S)=\sum_{i \in S \cap M} v(\{i\} \cup (\bigcup_{k=1} T_{ik})). \tag{2.26}$$

PROOF: For each $j \in T_{ik}$, there is an f_j such that $Y_{i(k,f_j)j}=1$, which implies $h_{kj} \geqq a_i(k,f_j)$. It follo from (2.4) that

$$\sum_{j \in T_{ik}} a_i(k,f_j) \geqq \sum_{f=0}^{|T_{ik}|-1} a_i(k, w_k^i-f). \tag{2.27}$$

Then we have

$$v\Big(\{i\} \cup \{\bigcup_{k=1}^{s} T_{ik}\Big)\Big)=\sum_{k=1}^{s}\Big(\sum_{j \in T_{ik}} h_{kj}+\sum_{f=1}^{w_k^i-|T_{ik}|} a_i(k,f)\Big).$$

By (2.23) and (2.27), we have

$$\begin{aligned} v(S) &= v^*(S^*)+\sum_{i \in S \cap M}\sum_{k=1}^{s}\sum_{f=1}^{w_k^i} a_i(k,f) \\ &= \sum_{Y_{i(k,f)j}=1}(h_{kj}-a_i(k,f))+\sum_{i \in S \cap M}\sum_{k=1}^{s}\sum_{f=1}^{w_k^i} a_i(k,f) \\ &\leqq \sum_{i \in S \cap M}\sum_{k=1}^{s}\Big(\sum_{j \in T_{ik}} h_{kj}+\sum_{f=1}^{w_k^i} a_i(k,f)-\sum_{f=0}^{|T_{ik}|-1} a_i(k, w_k^i-f)\Big) \\ &= \sum_{i \in S \cap M}\sum_{k=1}^{s}\Big(\sum_{j \in T_{ik}} h_{kj}+\sum_{f=1}^{w_k^i-|T_{ik}|} a_i(k,f)\Big) \\ &= \sum_{i \in S \cap M} v(\{i\} \cup \Big(\bigcup_{k=1}^{s} T_{ik}\Big)\Big). \end{aligned}$$

Since the characteristic function v satisfies superadditivity, we have (2.26). Q.E.

E CORE AND COMPETITIVE IMPUTATIONS

g be the function of A^* which specifies (u, v) for each $(u^*, v^*) \in A^*$ such that

$$u_i = \sum_{k=1}^{s} \sum_{f=1}^{w_k^i} (u^*_{i(k,f)} + a_i(k, f)) \qquad \text{for } \forall\, i \in M$$

$$v_j = v_j^* \qquad \text{for } \forall\, j \in N.$$

ine A by $A = g(A^*)$. We get the following theorem.

IEOREM I: The set of all competitive imputations of the market (M, N) is A.

OOF: Initially we show that every competitive imputation belongs to A. Let (u, v) be a competi-putation, and let (p, x) be a competitive equilibrium giving (u, v). For this (u, v) we define $(u^*,$

$$u^*_{i(k,f)} = \begin{cases} p_k - a_i(k, f) & \text{if } x_k^i < f \leqq w_k^i \\ 0 & \text{if } f \leqq x_k^i \end{cases}$$

$$v_j^* = v_j \qquad \text{for } \forall j \in N.$$

sily verified that $g(u^*, v^*) = (u, v)$. Hence we have to show that $(u^*, v^*) \in A^*$.

ce (u, v) is a competitive imputation, we have $u_t^* \geqq 0$ for $\forall\, t \in M^*$ and $v_j^* \geqq 0$ for $\forall j \in N$.

e suppose that there are $i(k, f) \in M^*$ and $j \in N$ such that

$$u^*_{i(k,f)} + v_j^* < b_{i(k,f)j}.$$

$f \leqq w_k^i$, then $u^*_{i(k,f)} = p_k - a_i(k, f)$, which implies

$$v_j = v_j^* < h_{kj} - a_i(k, f) - u^*_{i(k,f)} = h_{kj} - p_k.$$

ntradicts that (u, v) is a competitive imputation. If $f \leqq x_k^i$, then $u^*_{i(k,f)} = 0$ and $p_k \leqq a_i(k, f)$, mplies

$$v_j = v_j^* < h_{kj} - a_i(k, f) \leqq h_{kj} - p_k.$$

a contradiction. Hence we have

$$u_t^* + v_j^* \geqq b_{tj} \qquad \text{for } \forall\, t \in M^*,\ j \in N.$$

$u^*, v^*)$ is a feasible solution of (2.15).

ce (u, v) is a competitive imputation, by Lemma 1, we have

$$\sum_{i \in M} u_i + \sum_{j \in N} v_j = v(M \cup N).$$

Then, by (3.2) and (2.23), we have

$$\sum_{t \in M^*} u_t^* + \sum_{j \in N} v_j^* = \sum_{i \in M} u_i + \sum_{j \in N} v_j - \sum_{i \in M} \sum_{k=1}^{s} \sum_{f=1}^{w_k^i} a_i(k, f)$$

$$= v(M \cup N) - \sum_{i \in M} \sum_{k=1}^{s} \sum_{f=1}^{w_k^i} a_i(k, f) = v^*(M^* \cup$$

Hence we have $(u^*, v^*) \in A^*$ by the duality theorem of L.P. (see [1, Chapter 1]).

Next we show that every $(u, v) \in A$ is a competitive imputation. Let X be an extreme point solut of (2.14) for $M^* \cup N$, and let Y be given by (2.24). We define $x = (x^1, \ldots, x^m, x^{1'}, \ldots, x^{n'})$ by

$$x_k^i = w_k^i - \sum_{f=1}^{w_k^i} \sum_{j \in N} Y_{i(k,f)j} \qquad \text{for } \forall i \in M,\ k = 1, \ldots, s \tag{3.3}$$

$$x_k^j = \sum_{i \in M} \sum_{f=1}^{w_k^i} Y_{i(k,f)j} \qquad \text{for } \forall j \in N,\ k = 1, \ldots, s.$$

Now we suppose that there are $i_1(k, f_1), i_2(k, f_2) \in M^*$ and $j_1, j_2 \in N$ such that

$$u^*_{i_1(k,f_1)} + a_{i_1}(k, f_1) > u^*_{i_2(k,f_2)} + a_{i_2}(k, f_2),\ Y_{i_1(k,f_1)j_1} = Y_{i_2(k,f_2)j_2} = 1.$$

By the equilibrium theorem of L.P. (see [1, Chapter 1]), we have

$$u^*_{i_1(k,f_1)} + v^*_{j_1} = h_{kj_1} - a_{i_1}(k, f_1).$$

It follows that

$$u^*_{i_2(k,f_2)} + v^*_{j_1} < u^*_{i_2(k,f_1)} + a_{i_1}(k, f) - a_{i_2}(k, f_2) + v^*_{j_1} = h_{kj_1} - a_{i_2}(k, f_2) = b_{i_2(k,f_2)j_1}.$$

This contradicts that $(u^*, v^*) \in A^*$. Hence we have shown that

$$u^*_{i(k,f)} + a_i(k, f) = \text{constant} \qquad \text{for } \forall i(k, f) \in M^* \text{ such that } \sum_{j \in N} Y_{i(k,f)j} = 1. \tag{3.4}$$

By this, we can define $p = (p_1, \ldots, p_s)$ by

$$p_k = \begin{cases} u^*_{i(k,f)} + a_i(k, f) & \text{if } \exists\, i(k, f) \in M^* \colon \sum_{j \in N} Y_{i(k,f)j} = 1 \\ \min\{a_i(k, w_k^i) \mid i \in M\} & \text{if } \forall i(k, f) \in M^* \colon \sum_{j \in N} Y_{i(k,f)j} = 0. \end{cases}$$

By the equilibrium theorem of L.P., we have $u^*_{i(k,f)} = 0$ for $\forall\, i(k, f)$ such that $\sum_{\in N} Y_{i(k,f)j} = 0$. Hen we have, by using (2.4),

$$u_i = \sum_{k=1}^{s} \sum_{f=1}^{w_k^i} (u^*_{i(k,f)} + a_i(k,f))$$

$$= \sum_{k=1}^{s} \left(\sum_{f=1}^{x_k^i} a_i(k,f) + p_k(w_k^i - x_k^i) \right) \qquad \text{for } \forall\, i \in M.$$

$u^*_{i(k,f)} + v_j^* = h_{kj} - a_i(k,f)$ for $\forall\ i(k,f), j$ such that $Y_{i(k,f)j} = 1$, we have

$$v_j = \begin{cases} h_{kj} - p_k & \text{if } x^j = e^k \\ 0 & \text{if } x^j = 0. \end{cases}$$

fore we have to show that the pair (p, x) is a competitive equilibrium.

y the definition of x, it is clear that $\sum_{i \in N} x^i + \sum_{j \in N} x^j = \sum_{i \in M} w^i$. It is sufficient to show (2.19). Sup-

hat there is a $y_k^i > x_k^i$ such that

$$\sum_{f=1}^{y_k^i} a_i(k,f) + p_k(w_k^i - y_k^i) > \sum_{f=1}^{x_k^i} a_i(k,f) + p_k(w_k^i - x_k^i).$$

mplies

$$\sum_{f=x_k^{i+1}}^{y_k^i} a_i(k,f) > p_k(y_k^i - x_k^i).$$

4), we have $a_i(k, x_k^i + 1) > p_k$. By the definition of x, there is an f_o such that $a_i(k,f_0) \geqq a_i(k, x_k^i + 1)$

$\sum_{i \in N} Y_{i(k,f_0)j} = 1$, which implies $p_k = u^*_{i(k,f_0)} + a_i(k, f_0)$. Hence we have $u^*_{i(k,f_0)} < 0$, which is a

adiction. Suppose that there is a $y_k^i < x_k^i$ such that (3.5) holds for y_k^i. Then we have

$$p_k(x_k^i - y_k^i) > \sum_{f=y_k^{i+1}}^{x_k^i} a_i(k,f),$$

h implies $p_k > a_i(k, x_k^i)$. By the definition of x, there is an f_0 such that $a_i(k,f_0) \leqq a_i(k, x_k^i)$ and

$Y_{i(k,f_0)j} = 0$, which implies $u^*_{i(k,f_0)} = 0$. Since $p_k > a_i(k, f_0)$, there are $i_1(k, f_1) \in M^i$ and $j \in N$

that $Y_{i_1(k,f_1)j} = 1$. Since $u^*_{i_1(k,f_1)} + v_j^* = h_{kj} - a_{i_1}(k,f_1)$ and $p_k = a_{i_1}(k,f_1) + u^*_{i_1(k,f_1)}$, we have

$$u^*_{i(k,f_0)} + v_j^* = h_{kj} - a_{i_1}(k,f_1) - u^*_{i_1}(k,f_1) = h_{kj} - p_k < h_{kj} - a_i(k,f_0) \leqq b_{i(k,f_0)j}.$$

is a contradiction.

Suppose that there is a $j \in N$ such that $v_j = v_j^* < h_{kj} - p_k$. If $p_k = \min \{a_i(k, w_k^i) \mid i \in$ $= a_i(k, w_k^i)$, then we have $\sum Y_{i(k,w_k^i)j} = 0$, which implies $u^*_{i(k,w_k^i)} = 0$. Hence we have

$$u^*_{i(k,w_k^i)} + v_j^* = v_j^* < h_{kj} - p_k = h_{kj} - a_i(k, w_k^i) \leqq b_{i(k,w_k^i)j}.$$

This is a contradiction. If there is an $i(k, f) \in M^*$ such that $p_k = u^*_{i(k,f)} + a_i(k, f)$, then we have

$$v_j^* < h_{kj} - p_k = h_{kj} - u^*_{i(k,f)} - a_i(k, f),$$

which implies

$$u^*_{i(k,f)} + v_j^* < h_{kj} - a_i(k, f) \leqq b_{i(k,f)j}.$$

This is a contradiction. Q.E

Since A^* is nonempty, Theorem I means that the set of all competitive equilibria and the core nonempty. This fact may depend on the assumption that buyers want to buy at most one unit of goods.†

In the market (M^*, N), A^* coincides with not only the set of competitive imputations but also core. In the market (M, N), A does not necessarily coincide with the core, which is shown by Example But under a certain condition, A coincides with the core.

EXAMPLE 2: Let us consider the economy of Example 1. Then $A = \{(8+p_2,\ 14+p_1,\ 15-$ $15-p_1) \mid 4 \leqq p_1 \leqq 5,\ 2 \leqq p_2 \leqq 3\}$. A is the set of all competitive imputations of the economy by Th orem I. The imputation (12, 19, 11, 10) is not in A, but in the core, which is easily verified.

LEMMA 3: Let (u, v) be an arbitrary imputation in the core of the market (M, N). Let X be arbitrary extreme point solution of (2.14) for $M^* \cup N$, and let Y be given by (2.24). If there are at lea two sellers $i' \in M$ such that

$$a_{i'}(k, w_k^{i'}) \leqq \max \{a_i(k, f) \mid i \in M, \sum_{j \in N} Y_{i(k,f)j} = 1\}, \tag{3.6}$$

then there is a p_k such that

$$v_j = h_{kj} - p_k \qquad \text{for } \forall j \in N: \sum_{i \in M^*} Y_{ij} = 1. \tag{3.7}$$

PROOF: For $((T_{ik})_{k=1}^s)_{i \in M}$ given by (2.25), we have

$$v(M \cup N) = \sum_{i \in M} v\left(\{i\} \cup \left(\bigcup_{k=1}^{s} T_{ik}\right)\right).$$

† Shapley and Scarf [3] showed that there is a nonempty core in an economy where only indivisible goods are exchang and each trader wants to trade exactly one unit. But they gave an example of a coreless economy where each trader wants trade more than one unit.

e superadditivity of v, we have

$$u_i+\sum_{k=1}^{s}\sum_{i\in T_{ik}} v_j=v\Big(\{i\}\cup\Big(\bigcup_{k=1}^{s} T_{ik}\Big)\Big) \qquad \text{for } \forall i\in M.$$

we suppose that there are $j_1, j_2\in N$ such that

$$h_{k_0j_1}-v_{j_1}>h_{k_0j_2}-v_{j_2}$$

$$\sum_{i\in M}\sum_{f=1}^{w^i_{k_0}} Y_{i(k_0,f)j_1}=\sum_{i\in M}\sum_{f=1}^{w^i_{k_0}} Y_{i(k_0,f)j_2}=1.$$

re are i_1, $i_2(i_1\neq i_2)$ such that $j_1\in T_{i_1k_0}$ and $j_2\in T_{i_2k_0}$, then we have

$$u_{i_2}+\sum_{k=1}^{s}\sum_{j\in T_{i_2k}} v_j-v_{j_2}+v_{j_1}=v(\{i_2\}\cup(\bigcup_{k=1}^{s} T_{i_2k}))-v_{j_2}+v_{j_1}$$

$$=\sum_{k=1}^{s}\Big(\sum_{j\in T_{i_2k}} h_{kj}+\sum_{f=1}^{w^{i_2}_k-|T_{i_2k}|} a_{i_2}(k,f)\Big)-v_{j_2}+v_{j_1}$$

$$<\sum_{k=1}^{s}\Big(\sum_{j\in T_{i_2k}} h_{kj}+\sum_{f=1}^{w^{i_2}_k-|T_{i_2k}|} a_{i_2}(k,f)\Big)-h_{k_0j_2}+h_{k_0j_1}$$

$$\leqq v(\{i_2\}\cup(\bigcup_{k=1}^{s} T_{i_2k})\cup\{j_1\}-\{j_2\}).$$

is a contradiction. Next we consider the case where there is a unique i_1 for k_0 such that $|T_{i_1k_0}|>0$. re is an $i_1(k_0,f_0)$ such that $\sum_{j\in N} Y_{i_1(k_0,f_0)j}=1$ and $h_{k_0j_2}-v_{j_2}<a_{i_1}(k_0,f_0)$, then we have

$$u_{i_1}+\sum_{k=1}^{s}\sum_{j\in T_{i_1k}} v_j-v_{j_2}=\sum_{k=1}^{s}\Big(\sum_{j\in T_{i_1k}} h_{kj}+\sum_{f=1}^{w^{i_1}_k-|T_{i_1k}|} a_{i_1}(k,f)\Big)-v_{j_2}$$

$$<\sum_{k=1}^{s}\Big(\sum_{j\in T_{i_1k}} h_{kj}+\sum_{f=1}^{w^{i_1}_k-|T_{i_1k}|} a_{i_1}(k,f)\Big)-h_{k_0j_2}+a_{i_1}(k_0,f_0)$$

$$\leqq v(\{i_1\}\cup(\bigcup_{k=1}^{s} T_{i_1k})-\{j_2\}).$$

is a contradiction. Hence we have $h_{k_0j_2}-v_{j_2}\geqq a_{i_1}(k_0,f)$ for $\forall\, i_1(k_0,f)$ such that $\sum_{j\in N} Y_{i_1(k_0,f)j}=1$. ie assumption of the lemma, there is an $i_2(i_2\neq i_1)$ such that $a_{i_2}(k_0,w^{i_2}_{k_0})\leqq h_{k_0j_2}-v_{j_2}<h_{k_0j_1}-v_{j_1}$. $T_{i_2k_0}=\emptyset$, we have

$$u_{i_2}+\sum_{k=1}^{s}\sum_{j\in T_{i_2k}} v_j+v_{j_1}=\sum_{k=1}^{s}\left(\sum_{j\in T_{i_2k}} h_{kj}+\sum_{f=1}^{w_k^{i_2}-|T_{i_2k}|} a_{i_2}(k,f)\right)+v_{j_1}$$

$$<\sum_{k=1}^{s}\left(\sum_{j\in T_{i_2k}} h_{kj}+\sum_{f=1}^{w_k^{i_2}-|T_{i_2k}|} a_{i_2}(k,f)\right)-a_{i_2}(k_0,w_{k_0}^{i_2})+h_{k_0j_1}$$

$$=v(\{i_2\}\cup(\bigcup_{k=1}^{s}T_{i_2k})\cup\{j_1\}).$$

This is a contradiction. Therefore we have shown (3.7). Q

In Lemma 3, it was shown that the core has the property that if there are at least two seller a good who satisfy (3.6), then the good is exchanged at a common price. If there are at least two se for all goods who satisfy (3.6), then all goods are exchanged at common prices. Then the core coinc with the set of all competitive imputations, which will be shown by the following theorem.

THEOREM II: Assume that there are at least two sellers satisfying (3.6) for $\forall k$ such that

$$\sum_{i\in M}\sum_{f=1}^{w_k^i}\sum_{j\in N} Y_{i(k,f)j}\geqq 1.$$

Then the core of the market (M, N) coincides with A.

PROOF: Since A is the set of all competitive imputations by Theorem I and all competitive in tations belong to the core by Lemma 1, we have to show that the core is included by A.

Let (u, v) be an imputation in the core. By Lemma 3, we can define $p=(p_1, \ldots, p_s)$ by

$$(3.8)\qquad p_k=\begin{cases} h_{kj}-v_j & \text{if } \exists\, j\in N:\sum_{i\in M}\sum_{f=1}^{w_k^i} Y_{i(k,f)}=1 \\ \min\{a_i(k,w_k^i)\mid i\in M\} & \text{if } \forall\, j\in N:\sum_{i\in M}\sum_{f=1}^{w_k^i} Y_{i(k,f)j}=0. \end{cases}$$

We define (u^*, v^*) by

$$(3.9)\qquad u^*_{i(k,f)}=\begin{cases} p_k-a_i(k,f) & \text{if } \sum_{j\in N} Y_{i(k,f)j}=1 \\ 0 & \text{if } \sum_{j\in N} Y_{i(k,f)j}=0 \end{cases}$$

$$v_j^*=v_j \qquad \text{for } \forall\, j\in N.$$

It is easily verified that $g(u^*, v^*)=(u, v)$. Hence it is sufficient to show that $(u^*, v^*)\in A^*$. It is c that $v_j{}^*=v_j\geqq v(\{j\})=0$ *for* $\forall\, j\in N$. Suppose that there is an $i(k_0, f_0)$ such that $u_i^*{}_{(k_0,f_0)}<0$, $p_{k_0}<a_i(k_0,f_0)$. This means that there is a j_0 such that $Y_{i(k_0,f_0)j_0}=1$. Then we have

$$u_i + \sum_{k=1}^{s} \sum_{j \in T_{ik}} v_j - v_{j_0} = \sum_{k=1}^{s} \Big(\sum_{j \in T_{ik}} h_{kj} + \sum_{f=1}^{w_k^i - |T_{ik}|} a_i(k,f) \Big) - (h_{k_0 j_0} - p_{k_0})$$

$$< \sum_{k=1}^{s} \Big(\sum_{j \in T_{ik}} h_{kj} + \sum_{f=1}^{w_k^i - |T_{ik}|} a_i(k,f) \Big) - h_{k_0 j_0} + a_i(k_0, f_0)$$

$$\leqq v\Big(\{i\} \cup \Big(\bigcup_{k=1}^{s} T_{ik} \Big) - \{j_0\} \Big).$$

s a contradiction. Hence we have $u_t^* \geqq 0$ for $\forall t \in M^*$. It is easily verified that $v^*(M^* \cup N) = u_t^* + \sum_{j \in N} v_j^*$. Therefore we have to show that

$$u_t^* + v_j^* \geqq b_{tj} \qquad \text{for } \forall\, t \in M^*, j \in N.$$

ote that $a_i(k, f) \leqq a_{i'}(k, f')$ for $i(k, f), i'(k, f')$ such that $\sum_{j \in N} Y_{i(k,f)j} = 1$ and $\sum_{j \in N} Y_{i'(k,f')j} = 0$, se Y gives an optimal assignment. Suppose that there is an $i_0(k_0, f_0)$ such that $p_{k_0} > a_{i_0}(k_0, f_0)$ $\sum_{N} Y_{i_0(k_0,f_0)j} = 0$. By the assumption of the theorem, there is an i' $(i' \neq i_0)$ such that $a_{i_0}(k_0, f_0)$ $k_0, w_{k_0}^{i'})$. If $T_{i'k_0} \neq \emptyset$, then there is an $j_0 \in T_{i'k_0}$ such that $j_0 \notin \bigcup_{k=1}^{s} T_{i_0 k}$. If $T_{i'k_0} = \emptyset$, then there i'' such that $T_{i''k_0} \neq \emptyset$, because $p_{k_0} > a_{i_0}(k_0, f_0)$. Then it holds that $p_{k_0} > a_{i'}(k_0, w_{k_0}^{i'})$ and $\sum_{j \in N}$ $w_{k_0}^{i'})j = 0$. Hence we can assume that there is a j_0 such that $j_0 \in T_{ik_0}$ for some $i \neq i_0$. Then we

$$u_{i_0} + \sum_{k=1}^{s} \sum_{j \in T_{i_0 k}} v_j + v_{j_0} = \sum_{k=1}^{s} \Big(\sum_{j \in T_{i_0 k}} h_{kj} + \sum_{f=1}^{w_k^{i_0} - |T_{i_0 k}|} a_{i_0}(k,f) \Big) + (h_{k_0 j_0} - p_{k_0})$$

$$< \sum_{k=1}^{s} \Big(\sum_{j \in T_{i_0 k}} h_{kj} + \sum_{f=1}^{w_k^{i_0} - |T_{i_0 k}|} a_{i_0}(k,f) \Big) + h_{k_0 j_0} - a_{i_0}(k_0, f_0)$$

$$\leqq v(\{i_0\} \cup (\bigcup_{k=1}^{s} T_{i_0 k}) \cup \{j_0\}).$$

s a contradiction. Hence we have $p_k \leqq a_i(k, f)$ for $\forall\, i(k, f)$ such that $\sum_{j \in N} Y_{i(k,f)j} = 0$. ıppose that there are $i_0(k_0, f_0)$ and j_0 such that

$$u_{i_0}^*(k_0, f_0) + v_{j_0}^* < b_{i_0(k_0,f_0)j_0}.$$

Let $j_0 \in \bigcup_{k=1}^{s} T_{i_0 k}$. By the assumption of the theorem, there is an $i'(\neq i_0)$ such that $a_{i'}(k_0, w_{k_0}^{i'}) \leqq$ $\{a_i(k_0, f) \mid i \in M, \sum_{j \in N} Y_{i(k_0,f)j} = 1\}$. Since $p_{k_0} \geqq a_i(k_0, f)$ for $\forall\, i(k_0, f)$ such that $\sum_{j \in N} Y_{i(k_0,f)j} =$ have $u^*_{i'(k_0,\ w^{i'}_{k_0})} \leqq p_{k_0} - a_{i'}(k_0,\ w^{i'}_{k_0})$. Since $p_{k_0} \leqq a_{i_0}(k_0, f_0)$ if $\sum_{j \in N} Y_{i_0(k_0,f_0)j} = 0$, we have $u^*_{i_0}$ $\geqq p_{k_0} - a_{i_0}(k_0, f_0)$. It follows that

$$u^*_{i'(k_0, w^{i'}_{k_0})} + v^*_{j_0} \leqq p_{k_0} - a_{i'}(k_0, w^{i'}_{k_0}) + v^*_{j_0} < p_{k_0} - a_{1'}(k_0, w^{i'}_{k_0}) + h_{k_0 j_0} - a_{i_0}(k_0, f_0) - p_{k_0} + a_{i_0}(k_0, f_0$$

$$= h_{k_0 j_0} - a_{i'}(k_0, w^{i'}_{k_0}) \leqq b_{i'(k_0, w^{i'}_{k_0})\, j_0}.$$

This means that it is sufficient to consider the case where $j_0 \notin \bigcup_{k=1}^{s} T_{i_0 k}$. If $\sum_{j \in N} Y_{i_0(k_0,f_0)j} = 0$,

$u^*_{i_0(k_0,f_0)} = 0$. It follows that

$$u_{i_0} + \sum_{k=1}^{s} \sum_{j \in T_{i_0 k}} v_j + v_{j_0} = \sum_{k=1}^{s} \left(\sum_{j \in T_{i_0 k}} h_{kj} + \sum_{f=1}^{v^{i_0}_k - |T_{i_0 k}|} a_{i_0}(k, f) \right) + v^*_{j_0}$$

$$< \sum_{k=1} \left(\sum_{j \in T_{i_0 k}} h_{kj} + \sum^{w^{i_0}_k - |T_{i_0 k}|} a_{i_0}(k, f) \right) + h_{k l j_0} - a_{i_0}(k_0, f_0)$$

$$= v\left(\{i_0\} \cup \left(\bigcup_{k=1}^{s} T_{i_0 k} \right) \cup \{j_0\} \right).$$

This is a contradiction. If there is a j_1 such that $Y_{i_0(k_0,f_0)j_1} = 1$, then $u^*_{i_0(k_0,f_0)} = p_{k_0} - a_{i_0}(k_0, f_0)$ $v^*_{j_1} = h_{k_0 j_1} - p_{k_0}$. It follows that

$$-v_{j_1} + v_{j_0} < -h_{k_0 j_1} + p_{k_0} + b_{i_0(k_0,f_0)j_0} - u^*_{i_0(k_0,f_0)} = -h_{k_0 j_1} + h_{k_0 j_0}.$$

Then we have

$$u_{i_0} + \sum_{k=1}^{s} \sum_{j \in T_{i_0 k}} v_j - v_{j_1} + v_{j_0} < \sum_{k=1}^{s} \left(\sum_{j \in T_{i_0 k}} h_{kj} + \sum_{f=1}^{w^{i_0}_k - |T_{i_0 k}|} a_{i_0}(k, f) \right) - h_{k_0 j_1} + h_{k_0 j_0}$$

$$= v\left(\{i_0\} \cup \left(\bigcup_{k=1}^{s} T_{i_0 k} \right) \cup \{j_0\} - \{j_1\} \right),$$

which is a contradiction. Q.

xt we consider the core of the market (M, N) in the case where there is exactly one seller who s (3.6). For simplicity, we assume that in the market (M, N), one kind of good is exchanged, i.e., n the following, we omit all subscripts representing the number of goods. We get the following n.

IEOREM III: Assume that there is exactly one seller who satisfies (3.6). Let the seller be 1,

$$h_{1'} \geqq h_{2'} \geqq \ldots \geqq h_{n'}.$$

he core is the set of (u, v) such that

$$u_1 = \sum_{f=1}^{w^1-d} a_1(f) + \sum_{j=1'}^{d'} (h_j - v_j)$$

$$u_i = \sum_{f=1}^{w^i} a_i(f) \qquad \text{for } i = 2, \ldots, m,$$

$$a_1(w^1 - d + 1) \leqq h_j - v_j \leqq \min\,(a_2(w^2), \ldots, a_m(w^m)),$$

$$h_{d+1'} \leqq h_j - v_j \leqq h_j \qquad \text{for } j = 1', \ldots, d',$$

$$v_j = 0 \qquad \text{for } j = d+1', \ldots, n'.$$

$$= \sum_{f=1}^{w^1} \sum_{j \in N} Y_{1(f)j}.$$

OOF: It follows from Lemma 2 that an imputation (u, v) is in the core if and only if

$$u_i + \sum_{j \in T} v_j \geqq v(\{i\} \cup T) \qquad \text{for } \forall\, i \in M,\ T \subset N.$$

tially we show that (u, v) satisfying (3.12) and (3.13) is in the core. Since it is easily verified $, v)$ is an imputation, it is sufficient to show (3.14). Let $S = \{i\} \cup \{j_1, \ldots, j_r\}$, and let

$$h_{j_1} \geqq h_{j_2} \geqq \ldots \geqq h_{j_r}.$$

$) = \sum_{f=1}^{w^i-q} a_i(f) + \sum_{t=1}^{q} h_{j_t}$. Let $i \neq 1$. Then it is easily verified that $j_q \leqq d$ and $v_{j_t} \geqq h_{j_t} - a_i(w^i)$ $a_i(w^i - t + 1)$ for $\forall\, t (t \leqq q)$. It follows that

$$u_i + \sum_{t=1}^{r} v_j = \sum_{f=1}^{w^i} a_i(f) + \sum_{j_t \leqq d'} v_j$$

$$\geqq \sum_{f=1}^{w^i} a_i(f) + \sum_{t=1}^{q} v_{j_t} \geqq \sum_{f=1}^{w^i} a_i(f) + \sum_{t=1}^{q} (h_{j_t} - a_i(w^i - t + 1))$$

$$= \sum_{f=1}^{w^i-q} a_i(f) + \sum_{t=1}^{q} h_{j_t} = v(S).$$

Let $i = 1$. We put

$$T^1 = \{j_t \in T | j_t \leqq d'\} \text{ and } T^2 = \{j_t \in T | d' < j_t \leqq j_q\},$$

where $T = \{j_1, \ldots, j_r\}$. Note that $T^2 = \emptyset$ if $d' \geqq j_q$. Then, since $q = |T^1| + |T^2|$, we have $d - = d - q + |T^2|$. Since

$$h_{j_t} \leqq h_{d+1'} \leqq h_j - v_j \qquad \text{for } \forall j_t\,(j_t \geqq d + 1'), j(j \leqq d')$$

and

$$a_1(f) \leqq a_1(w^1 - d + 1) \leqq h_j - v_j \qquad \text{for } \forall f\,(f \geqq w^1 - d + 1), j(j \leqq d'),$$

we have

$$\sum_{\substack{j \notin T^1 \\ 1' \leqq j \leqq d'}} (h_j - v_j) \geqq \sum_{f = w^1 - d + 1}^{w^1 - q} a_1(f) + \sum_{j \in T^2} h_j.$$

Hence it follows that

$$u_1 + \sum_{t=1}^{r} v_{j_t} = \sum_{f=1}^{w^1-d} a_1(f) + \sum_{j=1'}^{d'} (h_j - v_j) + \sum_{j_t \leqq d'} v_{j_t}$$

$$= \sum_{f=1}^{w^1-d} a_1(f) + \sum_{j \in T^1} h_j + \sum_{\substack{j \notin T^1 \\ 1' \leqq j \leqq d'}} (h_j - v_j)$$

$$\geqq \sum_{f=1}^{w^1-d} a_1(f) + \sum_{j \in T^1} h_j + \sum_{f = w^1 - d + 1}^{w^1 - q} a_1(f) + \sum_{j \in T^2} h_j$$

$$= \sum_{f=1}^{w^1-q} a_1(q) + \sum_{t=1}^{q} h_{j_t} = v(S).$$

Next we show that any imputation in the core satisfies (3.12) and (3.14). By Lemma 2, we have

$$v(M \cup N) = v(\{1\} \cup \{1', \ldots, d'\}) + \sum_{i=2}^{m} v(\{i\}).$$

If (u, v) is an imputation in the core, then we have

$$u_1 + \sum_{j=1}^{d'} v_j = v(\{1\} \cup \{1', \ldots, d'\}) = \sum_{f=1}^{w^1-d} a_1(f) + \sum_{j=1'}^{d'} h_j$$

$$u_i = v(\{i\}) = \sum_{f=1}^{w^i} a_i(f) \qquad \text{for } i = 2, \ldots, m$$

$$v_j = 0 \qquad \text{for } j = d+1', \ldots, n'.$$

it is sufficient to show (3.13). Since $v_j \geqq v(\{j\})$ for $\forall\, j \in N$, we have $h_j - v_j \leqq h_j$ for $\forall\, j \in N$. If s a j_0 such that $1' \leqq j_0 \leqq d'$ and $a_1(w^1 - d + 1) > h_{j_0} - v_{j_0}$, then we have

$$\sum_{j'} v_j - v_{j_0} = \sum_{f=1}^{w^1-d} a_1(f) + \sum_{\substack{j=1' \\ j \neq j_0}}^{d'} h_j + h_{j_0} - v_{j_0} < \sum_{f=1}^{w^1-d} a_1(f) + \sum_{\substack{j=1 \\ j \neq j_0}}^{d'} h_j + a_1(w^1 - d + 1)$$

$$= v(\{1\} \cup \{1', \ldots, d'\} - \{j_0\}).$$

a contradiction. The other two inequalities can be proved without difficulty. We omit the proofs
n. Q.E.D.

he core of a market where there is exactly one seller satisfying (3.6) is quite different from that of et where there are at least two sellers satisfying (3.6). In the former, the seller may sell the goods ers at different prices (in Theorem III, seller 1 sells to buyer $j (j \leqq d')$ at prices $h_j - v_j$), but in er, sellers sell at common prices. It is a critical condition whether or not there are at least two satisfying (3.6).

REFERENCES

e, D., *The Theory of Linear Economic Models* (McGraw Hill, New York, 1960).

en, G., "On the Core of Linear Production Games," To appear in Mathematical Programing.

pley, L. and H. Scarf, "On Cores and Indivisibility," J. Math. Econ. *1*: 23–37 (1974).

pley, L. and M. Shubik, "The Assignment Game I: The Core," Intern. J. Game Theory *1*: 1–130 (1972).

uki, M. ed., *Developments in Game Theory* (Japanese), Tokyo-Tosho, Tokyo (1973).

ser, L. G., *Competition, Collusion and Game Theory* (Aldine, Atherton, 1972).

Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univer- y Press, Princeton, 1944; 2nd. ed. 1947; 3rd. ed. 1953).

'RIBUTION OF A LINEAR FUNCTION AND THE RATIO OF TWO EPENDENT LINEAR FUNCTIONS OF INDEPENDENT GENERALIZED GAMMA VARIABLES *

Henrick John Malik
University of Guelph
Guelph, Ontario

ABSTRACT

In this paper the exact distribution of a linear function and the ratio of two independent linear functions of independent generalized gamma variables is given.

TRODUCTION

et X be a random variable whose frequency function is

$$f(x;\, a,\, d,\, p) = (p(\Gamma(d/p)a^d)^{-1}x^{d-1}e^{-(x/a)^p},\ x > 0;\ a,\, d,\, p > 0.$$

1.1) is Stacy's [6] generalization of the gamma distribution. The familiar gamma, chi, chi-squared, ential and Weibull variates are special cases, as are certain functions of normal variates. Form also a function introduced by Amoroso [1] in analyzing the distribution of income. Stacy [6] udied some of the elementary properties of (1.1). Parr and Webster [5] have obtained expressions maximum likelihood estimators of the parameters of (1.1) and for their asymptotic variances and ances. There are several problems in physical sciences where one needs the distribution of a prod- notient, linear function, and the ratio of two independent linear functions of independent random es. Miller [4] mentions a number of such problems. Malik [2, 3] gives the exact distribution of otient and of the product of two independent generalized gamma variables. In this note the exact ution of a linear function and the ratio of two independent linear functions of independent gener- gamma variables is given.

esearch supported by National Research Council of Canada.

2. DISTRIBUTION OF A LINEAR FUNCTION OF INDEPENDENT GENERALIZED GAMMA VARIABLES

Let $X_1, X_2, \ldots, X_n$ be independently distributed with respective frequency functions given by

$$f(x_j;\, a_j,\, d_j,\, p_j) = [p_j \Gamma(d_j/p_j)\; a_j^{d_j}]^{-1} x_j^{d_j - 1} e^{-(x_j/a_j)^{p_j}}, \tag{2.1}$$

$$x_j > 0;\; a_j,\, d_j,\, p_j > 0 \qquad j = 1, 2, \ldots$$

Let the linear function be

$$Y = \sum_{j=1}^{n} \lambda_j X_j \tag{2.2}$$

where the X_j are defined as above. We may assume, without any loss of generality, that all the λ positive and greater than unity.

Now the characteristic function for the distribution of X_j is given by

$$\phi_j(t) = \sum_{r=0}^{\infty} \frac{\Gamma\left(\frac{d_j}{p_j} + \frac{r}{p_j}\right) a_j^r (it)^r}{r!\,\Gamma(d_j/p_j)}, \qquad j = 1, 2, \ldots, n. \tag{2.3}$$

Hence the characteristic function for the distribution of Y is given by

$$\phi_Y(t) = \prod_{j=1}^{n} \sum_{r=0}^{\infty} \frac{\Gamma\left(\frac{d_j}{p_j} + \frac{r}{p_j}\right) a_j^r (i\lambda_j t)^r}{r!\Gamma(d_j/p_j)} \tag{2.4}$$

$$= \sum_{r_1, r_2, \ldots, r_n}^{\infty} P_n(a_j, d_j, p_j, r_j)$$

where $P_n(a_j, d_j, p_j, r_j)$

$$= \prod_{j=1}^{n} \frac{\Gamma\left(\frac{d_j}{p_j} + \frac{r_j}{p_j}\right) a_j^{r_j} (i\lambda_j t)^{r_j}}{r_j!\,\Gamma(d_j/p_j)}.$$

Hence the distribution of Y is given by

$$f(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \phi_Y(t)\, dt \tag{2.5}$$

$$= \sum_{r_1, r_2, \ldots, r_n}^{\infty} P_n(a_j, d_j, p_j, r_j) \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity}\, (it)^{R_n} dt$$

where

$$R_n = r_1 + r_2 + \ldots + r_n.$$

t may be shown that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}(it)^{R_n}dt=\sum_{r=0}^{R_n}\binom{R_n}{r}(-1)^{R_n}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}(1-it)^{-[r-R_n]}dt.$$

ell known that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}(1-it)^{-r}dt=\frac{e^{-x}x^{r-1}}{\Gamma(r)}.$$

sing (2.5), (2.6), and (2.7), we have the exact distribution of a linear function $Y=\sum_{j=1}^{n}\lambda_j X_j$.

$$f(y)=\sum_{r_1,r_2,\ldots,r_n}^{\infty} P_n(a_j,d_j,p_j,r_j)\sum_{r=0}^{R_n}\binom{R_n}{r}(-1)^{r+R_n}\times\frac{e^{-y}(y)^{r-R_n-1}}{\Gamma[r-R_n]}.$$

STRIBUTION OF THE RATIO OF TWO INDEPENDENT LINEAR FUNCTIONS

et $U=\sum_{j=1}^{m}\lambda_j X_j/\sum_{j=1}^{n}\lambda'_j X'_j$ where the X_j and the X'_j are independent generalized gamma variables. all assume, without any loss of generality, that all the λ_j and the λ'_j are positive and greater than

or simplicity, let us write $U=Y/Z$, where

$\lambda_j X_j$ and $Z=\sum_{j=1}^{n}\lambda'_j X'_j$.

(2.8), the distributions of Y and Z are given by

$$f(y)=\sum_{r_1,r_2,\ldots,r_m}^{\infty} P_m(a_j,d_j,p_j,r_j)\sum_{r=0}^{R_m}\binom{R_m}{r}(-1)^{r+R_m}\times\frac{e^{-y}y^{r-R_m-1}}{\Gamma[r-R_m]},$$

$$f(z)=\sum_{r'_1,\ldots,r'_n=0}^{\infty} P_n(a_j,d_j,p_j,r'_j)\sum_{r'=0}^{R'_n}\binom{R'_n}{r'}(-1)^{r'+R'_n}\times\frac{e^{-z}z^{r'-R'_n-1}}{\Gamma[r'-R'_n]}.$$

all first derive the distribution of

$$W=\log U=\log Y-\log Z.$$

The characteristic function of $W=\log U$ is given by

$$\phi_W(t)=\int_0^\infty\int_0^\infty e^{iwt}f(y)f(z)\,dy\,dz \tag{3.3}$$

$$=\sum_{r_1,\ldots,r_m}^{\infty}P_m(a_j,d_j,p_j,r_j)\sum_{r=0}^{R_m}\binom{R_m}{r}(-1)^{r+R_m}$$

$$\times\sum_{r'_1,\ldots,r'_n=0}^{\infty}P_n(a_j,d_j,p_j,r'_j)\sum_{r'=0}^{R'_n}\binom{R'_n}{r'}(-1)^{r'+R'_n}$$

$$\times\sum_{r,\,r'=0}^{\infty}\frac{1}{\Gamma[r-R_m]\Gamma[r'-R'_n]}$$

$$\times\Gamma[r-R_m+it]\Gamma[r'-R'_n+it].$$

Hence the distribution of W is given by

$$f(w)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-iwt}\phi_W(t)\,dt. \tag{3.4}$$

Now note from (3.3) that in order to find $f(w)$, we need to calculate the following integral I:

$$I=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-iwt}\Gamma[r-R_m+it]\Gamma[r'-R'_n-it]dt. \tag{3.5}$$

If we make the transformation $r'=R'_n-it=-z$, (3.5) may be written as

$$I=e^{-w}[r'-R'_n]\,\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-wz}\Gamma[(r+r')-(R_m+R'_n)+z]. \tag{3.6}$$

Using a well known result from Whittaker and Watson [7], (3.6) may be written as

$$I=e^{-w[r'-R'_n]}[1+e^{-w}]^{-[(r+r')-(R_m+R'_n)]}\Gamma[(r+r')-(R_m+R'_n)]. \tag{3.7}$$

sing (3.3), (3.5), and (3.7) and making the transformation $U=e^{W}$, the distribution of the ratio ven by

$$f(u)=\sum_{r_1,\ldots,r_m}^{\infty} P_m(a_j, d_j, p_j, r_j)\sum_{r=0}^{R_m}\binom{R_m}{r}(-1)^{r+R_m}$$

$$\times\sum_{r'_1,\ldots,r'_n}^{\infty} P_n(a_j, d_j, p_j, r'_j)\sum_{r'=0}^{R'_n}\binom{R'_n}{r'}(-1)^{r'+R'_n}$$

$$\times\sum_{r,\,r'=0}^{\infty}\frac{u^{r-R_m-1}}{\beta(r-R_m, r'-R'_n)(1+u)^{[(r+r')-(R_m+R'_n)]}}.$$

REFERENCES

noroso, L., "Richerche Intorno Alla Curva dei Redditi," Ann. Mat. Pura Appl. Ser. *4 21:* 123–159 1925).

lik, H. J., "Exact Distribution of the Quotient of Independent Generalized Gamma Variables," Canad. Math. Bull. *10:* 463–465 (1967).

lik, H. J., "Exact Distribution of the Product of Independent Generalized Gamma Variables with the Same Shape Parameter," Ann. Math. Statist. *39:* 1751–1752 (1968).

ller, K. S., *Multi-dimensional Gaussian Distributions* (John Wiley & Sons, New York, 1964).

rr, V. B. and J. T. Webster, "A Method for Discriminating Between Failure Density Functions Used in Reliability Predictions," Technometrics *7:* 1–10 (1965).

cy, E. W., "A Generalization of the Gamma Distribution," Ann. Math. Statist. *33:* 1187–1192 1962).

ittaker, E. T. and G. N. Watson, *A Course of Modern Analysis* (Cambridge University Press, 920).

THE RELATIONSHIP BETWEEN THE FORCE RATIO AND THE INSTANTANEOUS CASUALTY-EXCHANGE RATIO FOR SOME LANCHESTER-TYPE MODELS OF WARFARE

James G. Taylor*

Department of Operations Research and Administrative Sciences
Naval Postgraduate School
Monterey, California

ABSTRACT

A "local" condition of winning (in the sense that the force ratio is changing to the advantage of one of the combatants) is shown to apply to all deterministic Lanchester-type models with two force-level variables. This condition involves the comparison of only the force ratio and the instantaneous force-change ratio. For no replacements and withdrawals, a combatant is winning "instantaneously" when the force ratio exceeds the differential casualty-exchange ratio. General outcome-prediction relations are developed from this "local" condition and applied to a nonlinear model for Helmbold-type combat between two homogeneous forces with superimposed effects of supporting fires not subject to attrition. Conditions under which the effects of the supporting fires "cancel out" are given.

TRODUCTION

nce the time of Lanchester (*see* [6]), analysts have employed simplified deterministic differential on models to obtain insights into the dynamics of combat (*see* [1], [2], [12]–[14]), even though com- tween two military forces is a complex random process (*see* Notes 1 and 2 of Taylor and Brown Today, Lanchester-type models of quite complex military systems have been developed which e a digital computer for their implementation (*see*, for example, Bonder and Honig [2]). Never-, a simple model of the combat process may yield an understanding of important relations t to perceive in a more complex model, and such insights may be useful in guiding higher resolu- mputerized investigations (*see* [1], [14]). For such simplified models (in particular, Lanchester- quations with two force-level variables), we will develop a simple, yet very basic, "local" condi- force superiority that sometimes allows prediction of battle outcome without explicitly solving † nchester-type combat equations (even when there are time-dependent attrition rate coefficients ng temporal variations in such factors as combatant posture, force separation, rates of target

is research was supported by the Office of Naval Research as part of the Foundation Research Program at the Naval luate School.

† work by Bonder and Farrel [1] and Taylor [10], [11] shows, in general the analytic (i.e., infinite series) solution to coefficient equations by itself provides little information about battle outcome because of its complexity.

acquisition and fire, etc.).* Such results are not only important in their own right but also useful i quantitative analysis of tactics (*see*, for example, Taylor [9]).

Recently, Taylor and Parry [12] studied the Riccati equation satisfied by the force ratio corresp ing to a linear† ordinary differential equation combat model with two force-level variables and sh how to determine battle outcome without explicitly solving the force-level equations. In this pape extend these results and show that a simple condition with a rich military interpretation explai these battle-outcome prediction results: namely, a force is "winning" when the force ratio exc the casualty-exchange ratio.** This "local" condition of force superiority applies to *all* Lanche type models with two force-level variables and sometimes yields sufficient conditions for victory when no Ricatti equation holds for the force ratio.

In this paper we give a completely general "local" condition for force superiority, and wit appropriate assumptions we develop "global" conditions for winning (i.e., conditions sufficie guarantee victory in a fixed force-ratio breakpoint battle). We then apply these general condition nonlinear infantry combat model with supporting fires not subject to attrition and obtain cond sufficient to guarantee victory without solving in detail. From studying this model we gain son sights into the effects of supporting fires: for example, when both sides' supporting fires are ec effective, they neutralize each other so that the battle's outcome is as though there were no supp fires (although the ending of battle is accelerated).

2. THEORETICAL DEVELOPMENTS

Let us consider combat between two homogeneous forces described by the following determi Lanchester-type equations for x, $y > 0$ [the first equation of (1) becomes, for example, $dx/dt =$ $x = 0$]

$$(1) \qquad \begin{cases} dx/dt = -F(t, x, y) & \text{with } x(t=0) = x_0, \\ dy/dt = -G(t, x, y) & \text{with } y(t=0) = y_0, \end{cases}$$

where $x(t)$ and $y(t)$ denote the X and Y force levels at time t, and F and G denote force-change (with a negative force-change rate signifying a net influx of replacements). When there are no re ments and withdrawals, F and G are simply casualty rates. To insure the existence of partial d tives needed in subsequent analysis, we assume that F and G are differentiable.

The *instantaneous* (or differential) *force-change ratio* dx/dy is given by

---

*In his well-known survey paper on the Lanchester theory of combat, Dolansky [4] suggested the development of ou predicting relations without solving in detail as one of several problems for future research. The work at hand is a step t resolution of this problem (*see* also Taylor and Parry [12]).

†We mean here that the force levels appear linearly in the right-hand sides of the differential equations rather th some type of Lanchester "linear law" arises.

**This interpretation only holds for cases of no replacements and withdrawals or, more generally, when the rates of r ment and withdrawal are equal.

here are no replacements and withdrawals, dx/dy is the *instantaneous* (or differential) *casualty-ge ratio:* from X's standpoint, it is the cost of reducing the enemy force level a unit amount. dx/dy is time-invariant (i.e., $\partial(dx/dy)/\partial t \equiv 0$ for all $t \geq 0$), we will say that the Lanchester-type ns (1) are *quasi-autonomous*, since they may be transformed to an autonomous system (*see*, mple, p. 163 of Petrovski [7]) by a change of the time scale. Such Lanchester-type equations have, mple, been considered by Farrell [1] and Taylor [8] (*see* also Note 4 of Taylor and Brown [11]). roducing the *force ratio* $u = x/y$, we find after some straightforward manipulations that*

$$u - \frac{dx}{dy} = \frac{\dfrac{du}{dt}}{\left\{-\dfrac{1}{y}\dfrac{dy}{dt}\right\}}.$$

ume for simplicity that $dy/dt < 0$, with other cases being handled in a straightforward manner. $dy/dt < 0$, then du/dt and $u - dx/dy$ have the same sign.

nsider now a battle which terminates at the first time that either of two given "breakpoint" atios is reached. These "breakpoint" force ratios, denoted as u_X^f when X wins and u_Y^f when Y atisfy $0 \leq u_Y^f < u_0 = u(t=0) < u_X^f \leq +\infty$. Corresponding to a fiight until the annihilation of one the other is the case in which $u_Y^f = 0$ and $u_X^f = +\infty$. As pointed out by Taylor and Parry [12], it appropriate to say that "the course of battle is moving towards an X victory" when $du/dt > 0$ ply, that "X is winning"). Thus, for $dy/dt < 0$ we have by (3) the following "local" condition win:

$$u > \frac{dx}{dy} = Q(t, x, y).$$

ıy be considered as an "instantaneous" condition for X-force superiority. It is of much greater , however, to have "global" conditions for X to win. Although we have not succeeded in develop-h conditions in general, we will now give results for a special case of fairly wide applicability. some combat models of interest (*see* next section), $Q(t, x, y)$ is a homogeneous function of zero in the force-level variables x and y (*see* pp. 108–110 of Courant [3]). Then, as is well known, take dx/dy to depend only on t and the ratio x/y. When this is true, we will say that Condition lds.

Condition (HO): $Q(t, x, y) = q(t, u)$, where $u = x/y$.

e the following assumptions about $q(t, u)$:

$$u \begin{cases} < q(t, u) & \text{for } 0 < u < u_+, \\ > q(t, u) & \text{for } u > u_+, \end{cases}$$

* result (3) is the key result from which all subsequent developments in this paper follow.

(A2) $$\left.\frac{\partial q}{\partial u}\right|_{u=u_+} < 1,$$

where u_+ denotes the positive root of the equation

(5) $$E(t, u=u_+) = u_+ - q(t, u_+) = 0.$$

If $u_+(t)$ is a nonincreasing function of time, for X to win we require that (4) holds only at

THEOREM 1: Assume that Condition (HO) holds and that $u_+(t)$ is a nonincreasing func time. Then $u_0 > \left(\frac{dx}{dy}\right)_0$ is a sufficient condition for X to win a fixed force-ratio breakpoint battl

PROOF: If $u_0 = u(t=0) = x_0/y_0 > (dx/dy)_0$, then $dy/dt < 0$ implies that $du/dt > 0$ so increases for t near zero. Since $u_+(t)$ is nonincreasing and $u_+(t) < u_0 \leq u(t)$, it is clear by (A1) must win, since u is always increasing. *Q.E.D.* We now give a condition that is necessary and suf for $u_+(t)$ to be nonincreasing.

THEOREM 2: Assume that Condition (HO) holds. Then $u_+(t)$ is a nonincreasing function o if and only if $\frac{\partial q_+}{dt} = \left.\frac{\partial q}{\partial t}\right|_{u=u_+} \leq 0.$

PROOF: Differentiating the identity (5), we obtain

$$\frac{du_+}{dt} = \frac{\frac{\partial q_+}{\partial t}}{1 - \frac{\partial q_+}{\partial u}},$$

whence follows the theorem by (A2). Q.E.D. As an immediate corollary, we have

COROLLARY 2.1: Assume that Condition (HO) holds and that $\frac{\partial q_+}{\partial t} = \left.\frac{\partial q}{\partial t}\right|_{u=u_+} \leq 0$. Then $\left(\frac{dx}{dy}\right)_0$ is a sufficient condition for X to win a fixed force-ratio breakpoint battle.

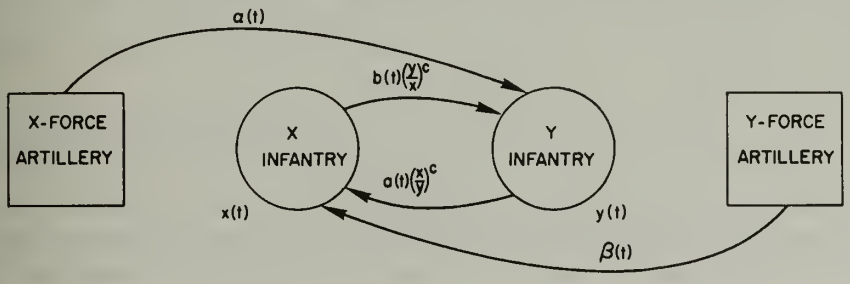
Also

COROLLARY 2.2: Assume that Condition (HO) holds for quasi-autonomous Lanchester Equations (1). Then X wins a fixed force-ratio breakpoint battle if and only if $u_0 > \left(\frac{dx}{dy}\right)_0$.

3. APPLICATION TO COMBAT WITH SUPPORTING FIRES

Let us consider combat between two heterogeneous forces, each composed of infantry a supporting weapon system (*see* Weiss [14]), denoted here as artillery. We assume that attrition bet the two infantries follows Helmbold's [5] modification of Lanchester's [6] equations of "modern fare" (*see* also [10], [11]) to account for inefficiencies of scale for the larger infantry when infa force sizes are grossly unequal. We may consider each infantry to deliver aimed fire against the en

y (*see* [12], [13]) but with the efficiency of its fire effectiveness modified by Helmbold's scheme at e force ratios. We assume that the supporting weapon systems are not subject to attrition and area fire against the enemy infantry. The supporting weapons' effects are consequently super-d on the Helmbold-type attrition process. The following Lanchester-type equations have been esized (*see* [5], [12], [13]) to describe this situation (*see* Figure 1):



1. Helmbold-type combat between two homogeneous forces (infantry) with supporting weapons (artillery) not subject to attrition.

$$\begin{cases} \dfrac{dx}{dt} = -a(t)\left(\dfrac{x}{y}\right)^c y - \beta(t)x & \text{with } x(t=0)=x_0, \\ \dfrac{dy}{dt} = -b(t)\left(\dfrac{y}{x}\right)^c x - \alpha(t)y & \text{with } y(t=0)=y_0. \end{cases}$$

ral variations in the (nonnegative) attrition-rate coefficients* are allowed in this model. The model (6) reduces to that studied by Taylor and Parry [12] when $c=0$.†

e see that for (6) dx/dy is homogeneous of degree zero in x and y [i.e. Condition (HO) holds for odel (6)] and

$$q(t, u) = u^{(1-d/2)}\left\{\frac{a(t)+\beta(t)u^{d/2}}{\alpha(t)+b(t)u^{d/2}}\right\},$$

$d=2(1-c)$. We assume that $d\epsilon[0,2]$. When c varies between 0 and 1, d varies between 2 and 0. serve that $dy/dt<0$. Let us now verify that (A1) and (A2) are satisfied. To show this we consider $=u-q(t, u)$, where $q(t, u)$ is given by (7). Then

$$E(t, u)=\{u^{(1-d/2)}/(b(t)u^{d/2}+\alpha(t))\}\cdot h(t, u),$$

$$h(t, u)=b(t)u^d+\{\alpha(t)-\beta(t)\}u^{d/2}-a(t).$$

e prediction of $a(t)$ and $b(t)$ (for $c=0$) from weapon system performance data is discussed in [1]. See [12] for modelling o $\alpha(t)$ and $\beta(t)$.

Helmbold [5] for a discussion of how different classic constant-coefficient homogeneous-force (i.e., $\alpha=\beta=0$) combat ions (i.e., square law, logarithmic law, etc.) correspond to various values of $c\epsilon[0, 1]$.

Considering (8), we see that, for $u > 0$, the zeros of $E(t, u)$ coincide with those of $h(t, u)$. A $u < q(t, u)$ if and only if $h(t, u) < 0$. Noting that $h(t, u)$ is a quadratic form in $u^{d/2}$ [i.e. the substitu $v = u^{d/2}$ yields $h(t, u = v^{2/d}) = h_1(t, v) = b(t)v^2 + \{\alpha(t) - \beta(t)\}v - a(t)]$, we readily verify that (A1 satisfied, since $h(t, u = 0) = -a(t) < 0$, for a given value of t the equation $h_1(t, v) = 0$ has a sir positive root (denoted as v_+), and $h_1(t, v)$ is positive for $v > v_+$. Recalling that $E(t, u) = u - q(t, u)$. see that (A2) is equivalent to $\partial E/\partial u|_{u=u_+} > 0$. Recalling that $h(t, u = u_+) = 0$, we obtain from differ tiating (8) that

$$\left.\frac{\partial E}{\partial u}\right|_{u=u_+} = \{u_+^{(1-d/2)}/(b(t)u_+^{d/2} + \alpha(t))\} \cdot \left.\frac{\partial h}{\partial u}\right|_{u=u_+}. \tag{10}$$

Considering (10), we see that, for $u > 0$, (A2) is satisfied if and only if $\partial h/\partial u|_{u=u_+} > 0$. From (9) obtain

$$\frac{\partial h}{\partial u} = \{u^{(d/2-1)} \cdot d\}\{2b(t)u^{d/2} + [\alpha(t) - \beta(t)]\}. \tag{11}$$

Using (14) below, we obtain from (11) that

$$\left.\frac{\partial h}{\partial u}\right|_{u=u_+} = \{u_+^{(d/2-1)} \cdot d\}\sqrt{(\beta(t) - \alpha(t))^2 + 4a(t)b(t)} > 0,$$

and we have shown that (A2) holds.

Recalling that $E(t, u) = u - q(t, u)$, where $q(t, u) = dx/dy$, our new general condition (4) of force superiority yields via (8) that $h(t, u) > 0$ is a "local" condition for X to win a fixed force-ra breakpoint battle. Considering (9), we may write this "instantaneous" condition of X-force superiority

$$b(t)x^d + \{\alpha(t) - \beta(t)\}x^{d/2}y^{d/2} > a(t)y^d, \tag{12}$$

which for equal effectiveness of the supporting units [i.e., $\alpha(t) = \beta(t)$] becomes the "instantaneou square law in the new variables $X = x^{d/2}$, $Y = y^{d/2}$

$$b(t)x^d > a(t)y^d. \tag{13}$$

When $\alpha(t) = \beta(t) = 0$, we see that $\partial q/\partial t \leqslant 0$ for $R(t) = a(t)/b(t)$ nonincreasing. In this case, holding at $t = 0$ is a sufficient condition for X to win a fixed force-ratio breakpoint battle by Coroll 2.1. In general (for at least one of $\alpha(t)$ and $\beta(t)$ greater than zero), it is more convenient to obt from (5)

$$\{u_+(t)\}^{d/2} = \{\beta(t) - \alpha(t) + \sqrt{(\beta(t) - \alpha(t))^2 + 4a(t)b(t)}\}/\{2b(t)\}. \tag{14}$$

$(t) = a(t)/b(t)$ and $S(t) = \{\beta(t) - \alpha(t)\}/\sqrt{a(t)b(t)}$. Then $R(t)$ and $S(t)$* being nonincreasing ufficient condition for $u_+(t)$ to be nonincreasing. Thus, (12) holding at $t=0$ is sufficient for an ory when $R(t)$ and $S(t)$ are nonincreasing by Theorem 1. Furthermore, since Condition (HO) for the Lanchester-type Equations (6) (i.e., dx/dy is a homogeneous function of degree zero in rce-level variables x and y), we know by Theorem 2 that $u_+(t)$ nonincreasing implies that $dy)/\partial t|_{u=u_+} \leqq 0$ and conversely. For constant coefficients, a necessary and sufficient condition for X able to win a fixed force-ratio breakpoint battle is that $bx_0^d + (\alpha-\beta)x_0^{d/2}y_0^{d/2} > ay_0^d$ by Corollary nce Condition (HO) holds and by (8) and (9) $u_0 > (dx/dy)_0$ if and only if $bx_0^d + (\alpha-\beta)x_0^{d/2}y_0^{d/2} > ay_0^d$.

MMARY AND CONCLUSIONS

very military man intuitively knows that the force ratio and the (instantaneous) casualty-exchange fluence the outcome of battle. In this paper we have shown that these two ratios may be quanti- y related for Lanchester-type models of warfare to develop outcome-predicting relations without in detail. Moreover, even the exact functional form of the combat equations need not be known dict battle outcome if one knows how the force-change ratio will change over the course of battle. e have shown how the force ratio and the instantaneous force-change ratio may be used to ine whether the force ratio is becoming more or less favorable to one of the combatants. General e-predicting relations were developed for a fixed force-ratio breakpoint battle (a special case ch is a "fight to the finish") and then applied to a nonlinear combat model for which an analytic n has not been obtained. Our outcome-prediction results here extend and generalize those of and Parry [12], who studied only particular models (all linear), and allow a uniform treatment ome prediction. For example, let us consider the Lanchester-type equations

$$dx/dt = -a(t)f(t, x, y), \qquad dy/dt = -b(t)f(t, x, y),$$

ch $a(t)$, $b(t)$, and $f(t, x, y) > 0$. This model (15) is readily handled by our results and yields the taneous" linear law $b(t)x > a(t)y$ for X to be winning.

summary, we state the following conclusions:

A general "local" condition of force superiority which applies to *all* deterministic Lanchester- odels with two force-level variables may be based on comparing the force ratio with the instan- s force-change ratio (both expressed as friendly to enemy).

) For no replacements and withdrawals, a side is winning "instantaneously" when the force xceeds the differential casualty-exchange ratio. Furthermore, if the instantaneous casualty- ge ratio decreases over time (as the force levels themselves decrease), then this condition need ld at $t = 0$ to predict victory in a fixed force-ratio breakpoint battle (a special case of which is t to the finish").

I) When the instantaneous force-change ratio is a homogeneous function of degree zero in the vel variables, particularly convenient outcome-predicting relations may be obtained.

---

se are the same parameters introduced by Taylor and Parry [12] for (6) with $c=0$: $R(t)$ represents the relative effective- o X) of the primary units, while $S(t)$ represents the net effectiveness of Y's supporting units normalized by the "intensity" t between the primary units.

entific verification of such models (as with any combat model) is still an unresolved question (*see* Note 1 of Taylor and 1]).

(IV) For Helmbold-type combat between two homogeneous forces with supporting fires subject to attrition, if the supporting fires are equally effective, their effects "cancel out" (i.e., t battle's outcome, although accelerated, is the same as though they were not present).

REFERENCES

[1] Bonder, S. and R. Farrell (Editors), "Development of Models for Defense Systems Plannin Report No. *SRL 2147 TR 70–2 (U)* (Systems Research Laboratory, The University of Michig Ann Arbor, Michigan, Sept. 1970).

[2] Bonder, S. and J. Honig, "An Analytic Model of Ground Combat: Design and Applicatio Proceedings U.S. Army Operations Research Symposium *10:* 319–394 (1971).

[3] Courant, R., *Differential and Integral Calculus, Volume II* (Interscience Publishers, Inc., N York, 1936).

[4] Dolansky, L., "Present State of the Lanchester Theory of Combat," Operations Research 344–358 (1964).

[5] Helmbold, R., "A Modification of Lanchester's Equations," Operations Research *13:* 857–8 (1965).

[6] Lanchester, F. W., "Aircraft in Warfare: The Dawn of the Fourth Arm—No. V., The Princi of Concentration," Engineering *98:* 422–423 (1914) (reprinted on pp. 2138–2148 of the *Wo of Mathematics, Vol. IV,* J. Newman (Editor), Simon and Schuster, New York, 1956).

[7] Petrovski, I., *Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, New Jersey, 19 (reprinted by Dover Publications, Inc., New York, 1973).

[8] Taylor, J., "A Note on the Solution to Lanchester-Type Equations with Variable Coefficient Operations Research *19:* 709–712 (1971).

[9] Taylor, J., "Lanchester-Type Models of Warfare and Optimal Control," Nav. Res. Log. Quart. 79–106 (1974).

[10] Taylor, J., "Solving Lanchester-Type Equations for 'Modern Warfare' with Variable Coefficient Operations Research *22:* 756–770 (1974).

[11] Taylor, J. and G. Brown, "Canonical Methods in the Solution of Variable-Coefficient Lanches Type Equations of Modern Warfare," Operations Research *24,* 44–69 (1976).

[12] Taylor, J. and S. Parry, "Force-Ratio Considerations for Some Lanchester-Type Models of W fare," Operations Research *23:* 522–533 (1975).

[13] Weiss, H., "Lanchester-Type Models of Warfare," Proc. First International Conference on Ope tional Research: 82–98 (John Wiley, New York, 1957).

[14] Weiss, H., "Some Differential Games of Tactical Interest and the Value of a Supporting Weap System," Operations Research *7:* 180–196 (1959).

U. S. GOVERNMENT PRINTING OFFICE : 1976 O - 599-445

INFORMATION FOR CONTRIBUTORS

NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of information in logistics and will publish research and expository papers, including those in areas of mathematics, statistics, and economics, relevant to the over-all effort to improve ency and effectiveness of logistics operations.

nuscripts and other items for publication should be sent to The Managing Editor, NAVAL RCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. nuscript which is considered to be suitable material for the QUARTERLY is sent to one referees.

nuscripts submitted for publication should be typewritten, double-spaced, and the author etain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted original.

hort abstract (not over 400 words) should accompany each manuscript. This will appear ead of the published paper in the QUARTERLY.

re is no authorization for compensation to authors for papers which have been accepted ication. Authors will receive 250 reprints of their published papers.

ders are invited to submit to the Managing Editor items of general interest in the field ics, for possible publication in the NEWS AND MEMORANDA or NOTES sections UARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

JUNE 197
VOL. 23,
NAVSO P-

CONTENTS

ARTICLES

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217